2011-09

Free-text disease classification

Maxey, Craig.

Monterey, California. Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

FREE-TEXT DISEASE CLASSIFICATION

by

Craig Maxey

September 2011

Thesis Advisor:	Lyn R. Whitaker
Second Reader:	Samuel E. Buttrey

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
23–9–2011	Master's Thesis	2102-06-01—2104-10-31

4. TITLE AND SUBTITLE

Free-Text Disease Classification

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Craig Maxey

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Postgraduate School
Monterey, CA 93943

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Department of the Navy

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. **IRB Protocol Number: NPS.2011.0062-IR-EP5-A**

14. ABSTRACT

Modern medicine produces data with every patient interaction. While many data elements are easily captured and analyzed, the fundamental record of the patient/clinican interaction is captured in written, free-text. This thesis provides the foundation for the Military Health System to begin building an auto classifier for ICD9 diagnostic codes based on free-text clinican notes. Support Vector Machine models are fit to approximately 84,000 free-text records providing a means to predict ICD9 codes for other free-text records. While the research conducted in this thesis does not provide a consumate ICD9 classification model, it does provide the foundation required to further more detailed analysis.

15. SUBJECT TERMS

Statistical Learning, ICD9 Auto Classification, Medical Data Mining

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
Unclassified	Unclassified	Unclassified	UU	107	

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

FREE-TEXT DISEASE CLASSIFICATION

Craig Maxey
Lieutenant, United States Navy
B.A., The Citadel, 2004
Masters of Business Administration, University of Alabama at Birmingham, 2006
Masters of Healthcare Administration, University of Alabama at Birmingham, 2007

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 2011

Author: Craig Maxey

Approved by: Lyn R. Whitaker
Thesis Advisor

Samuel E. Buttrey
Second Reader

Robert F. Dell
Chair, Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Modern medicine produces data with every patient interaction. While many data elements are easily captured and analyzed, the fundamental record of the patient/clinican interaction is captured in written, free-text. This thesis provides the foundation for the Military Health System to begin building an auto classifier for ICD9 diagnostic codes based on free-text clinican notes. Support Vector Machine models are fit to approximately 84,000 free-text records providing a means to predict ICD9 codes for other free-text records. While the research conducted in this thesis does not provide a consumate ICD9 classification model, it does provide the foundation required to further more detailed analysis.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

AHLTA	Armed Forces Health Longitudinal Technology Application
CDM	Clinical Data Mart
CDM ID	Clinical Data Mart Patient ID
CHCS	Composite Health Care System
DOD	Department of Defense
FMP	Family Member Prefix
ICD9	International Disease Classification Code, 9th Revision
Idf	Inverse Document Frequency
MEPRS	Medical Expense and Reporting System
MHS	Military Health System
MTF	Military Treatment Facility
NMCP	Naval Medical Center Portsmouth
SVM	Support Vector Machine
TDM	Term Document Matrix
Tf	Term Frequency
TfIdfN	Normalized Inverse Document Frequency
XML	Extensible Markup Language

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

Coding is directly related to a physician's revenue stream and, as such, receives a great deal of attention in the private sector. Most physicians contract with external agencies to code diagnoses from each patient encounter to ensure that the physician receives maximum reimbursement. In contrast, the Military Heatlh System is presented with a different problem when it comes to coding. While the Military Health System does not receive payment for each patient visit, coding plays an important role in justifying the medical budget, ensuring leadership understands the trends of patient utilization and providing data for physician review. Unfortunately, subjectivity in coding mitigates analysis performed across the healthcare system. Lack of confidence in accuracy and consistency of coding leaks into the decision-making process. Leaders are able to reject analysis when it doesn't fit with their notions of how the system is performing. This thesis is the first to utilize statistical learning to provide the Military Health System with an objective and automated method to code patient diagnosis and should provide a foundation for further autoclassifcation research. The patient record data collected by the Military Health System allows a variety of methodologies to be implemented; however, this thesis will focus on a statistical learning method, support vector machines, for assigning diagnostic codes, specifically ICD9 codes, to records of patient visits based on the free-text physician notes found in those records.

Six months of data, which constituted 124,766 patient encounters, is collected from Naval Medical Center Portsouth. These encounters are randomized and split into training and test sets. The training set included 105,994 records and is split into a training set and a validation set. The training set consists of 84,919 records and the validation set consists of 21,075 records. These records include demographic data as well as the free-text note that will be the focus of this study.

The Term Document Matrix is the key data structure when developing a text classifier. Its construction involves analyzing the corpus of medical records and transforming each document into a vector of term counts or term frequencies (tf). An example of such a document vector is

$$\begin{bmatrix} \#\{Reviewed\} \\ \#\{Xray\} \\ \#\{Patient\} \\ \vdots \\ \#\{Infection\} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 3 \\ \vdots \\ 1 \end{bmatrix}$$

in which the term "Reviewed" appears once, "Xray" does not appear, and so on. These vectors are used to construct the Term Document Matrix whose elements are functions of the term frequencies. Each row of the Term Document Matrix represents a term from the corpus and each column represents a document. Specifically, the elements, x_{ij}, of a Term Document Matrix are associated with the occurence of term i, in document j. That is, x_{ij} might be taken to be the frequency of the i^{th} word in the j^{th} document. We analyze three Term Document Matrix configurations in this thesis.

Support Vector Machines (SVMs) are a supervised statistical learning technique introduced by Corinna Cortes and Vladimir Vapnik. The support vector machine seeks to find a hyperplane that maximizes the margin between two classes, say positive and negative, in a high-dimensional space. This hyperplane is known as the optimal hyperplane. For example, consider the two classes represented by orange plus signs, positive examples, and blue hyphens, negative examples, in Figure 1. The goal is to find the linear function of the two variables x_1 and x_2 which yields the widest separation, or largest margin, between the two classes. In Figure 1, the optimal seperating hyperplane is in fact a line. The margin is indicated by the dashed lines parallel to the optimal seperating hyperplane. The points on the margins are called the support vectors.

In Figure 1, the two classes are separable in two dimensions. Classes are not always separable in their original space. By mapping them to a sufficiently high-dimensional space, classes can be made seperable. SVM classification based on seperable classes without any crossover produces what is known as a hard margin classifier. In addition to hard margin classifiers, a soft margin classifier exists where the positive and negative examples are not necessarily required to be linearly seperable. This is accomplished using a penalty function described in more detail in this thesis.

To judge the performance of the SVM we utilize common text categorization metrics to evaluate

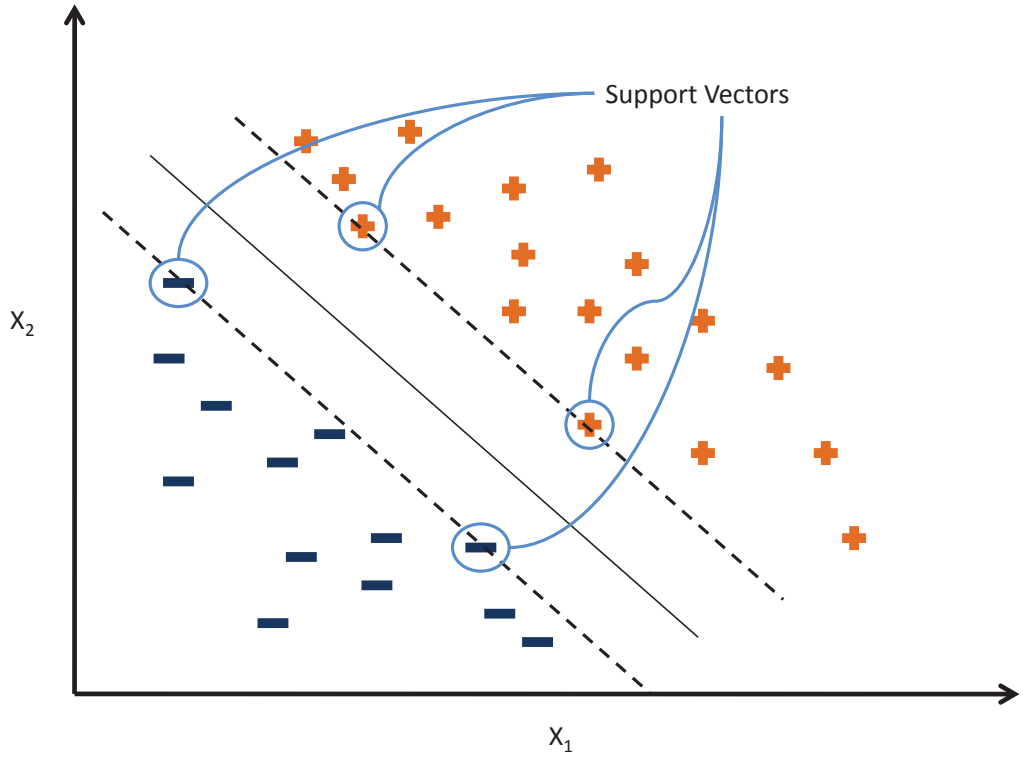


Figure 1: Example of a linearly separable problem in 2 dimensions. Support vectors define the margin of largest separation between the two classes.

the performance of the support vector machine models. These metrics include precision, a fraction that describes the confidence we have when the model predicts a positive outcome, recall, a fraction that describes how well the model identifies positive records out of the total number of positive records, and F-score, the harmonic mean of precision and recall providing insight into the overall performance of the support vector machine. The F-score provides the final measurement of success for this study.

Results indicate that the SVM alone does not provide a consumate classification technique for Military Health System free-text records. This thesis provides a foundation for future work in free-text disease classification specific to the Military Health System by providing a framework for data manipulation, and memory and analytically efficient data structures for analysis.

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgements

To my wife, without you this would have been impossible. Your faithful encouragement, understanding and love are the reason I have made it this far. Thank you for supporting my dreams, no matter how ridiculous, and for encouraging me to better myself. Now it is time to change my focus.

To my children, may this accomplishment serve as a reminder that you can accomplish anything you put your mind to. Nothing is impossible, but some things are very hard.

To my parents, thank you for love and support. If it were not for your instilling the value of education in me, I would never have made it this far. Not bad for a kid you didn't think would graduate high school, huh?

To my advisors, your hours of tutelage, advice and work are appreciated more than you know. Thank you for being educators, for reducing my ignorance and for showing me the joy of research. May your efforts be blessed and the culmination of your work be worthwhile.

And finally to my new friends. Thank you for your help and guidance over the past two years. I never would have made it through this program alone. My greatest gifts from NPS are a newfound knowledge and your friendship. Now. . . let's get back to the golf course!

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

Healthcare delivery in the United States has been under a great deal of scrutiny since as early as the 1920s when a study, commissioned by President Coolidge, addressed rising costs of healthcare delivery to Americans. Since that time, healthcare has been a lightning rod on the national stage on multiple occasions and is always an issue of public debate. The delivery of healthcare is constantly changing and rife for new and elegant ways to enhance the patient/provider experience while reducing costs.

At the turn of the 20th century healthcare practitioners were compensated with point of service bartering, entailing currency ranging from livestock to professional/manual services. [1] One hundred years later, third party payers and government agencies provide the financial infrastructure required to care for patients. As access to healthcare has expanded, the infrastructure to support the payment system has grown. With this growth, an entire industry of personnel to transcribe, interpret, bill, review and pay for services rendered to patients has emerged. The payment process hinges on the idea that diseases and procedures can be classified into groups that are homogenous across the medical spectrum. As early as the mid 1800s, clinicians were inventing simple ways to objectively communicate information about how patients fit into these homogeneous categories. One such schema for communication is the International Disease Classification coding system. Currently in its 9th revision, these codes have been adapted and grown throughout the years to their current form (ICD9), with over 7,000 code roots with numerous extenders for each code. Codes classify a disease so that a physician's diagnosis, regardless of semantics, can be understood and used throughout the world. Standardized codes are used for epidemiological study, medical research and billing throughout the medical community including the Military Health System (MHS). It is these codes, and how they are assigned, that are the focus of this study.

1.1 Problem Statement

Coding is directly related to a physician's revenue stream and, as such, receives a great deal of attention in the private sector. Most physicians contract with external agencies to code diagnoses from each patient encounter to ensure that the physician receives maximum reimbursement. In contrast, the MHS is presented with a different problem when it comes to coding. While the

MHS does not receive payment for each patient visit, coding plays an important role in justifying the medical budget, ensuring leadership understands the trends of patient utilization and providing data for physician review. Unfortunately, subjectivity in coding mitigates analysis performed across the healthcare system. Lack of confidence in accuracy and consistency of coding leaks into the decision-making process. Leaders are able to reject analysis when it doesn't fit with their notions of how the system is performing. This thesis is the first to utilize statistical learning to provide the MHS with an objective and automated method to code patient diagnosis and should provide a foundation for further autoclassifcation research. The patient record data collected by the MHS allows a variety of methodologies to be implemented; however, this thesis focuses on methods for assigning diagnostic codes, specifically ICD9 codes, to records of patient visits based on the free-text physician notes found in those records.

1.2 Statistical Learning and Text Classification

Statistical learning is a branch of statistics that promotes the study of algorithms which allow computers to improve their performance of a task given a training set of experiences. [2] As data sets grow larger, statistical learning has proven to be beneficial in science, finance and industry. Statistical learning models have been used to predict quantitative outcomes, such as stock prices, and categorical outcomes, such as the classification of email as spam or not spam, sucessfuly in many different industries. Medicine has an abundance of free-text data that is currently only useful to clinical professionals. These records, while rich in detail, are cumbersome and are not easily accessed, interpreted, nor profitable to practitioners. A solution to this problem may be found in the use of statistical text classification.

Text classification attempts to categorize large collections of texts, known as a corpus, into predefined categories using statistical and machine learning techniques. What makes text classification difficult is that each document can belong to multiple, exactly one, or no category at all. Furthermore, texts within each category can be written by different authors, contain different vocabularies, or have different semantic structures. Practitioners of text classification attempt to train models from data sets to automatically perform the categorization task.

Research shows that three distinct methodologies, naïve bayes, k-nearest neighbors and support vector machines, are often used to classify text. [3, 4, 5, 6] Each of these methodologies have distinct advantages that are highlighted by multiple authors. While each of the methodologies have proven succesful in previous research, support vector machines (SVM) appear to be the best solution for classifying diagnosis codes based on the text of the physician summaries [7],

which is the focus of this thesis. SVM's flexiblity, indifference to high dimensional sample spaces, and robustness to deal with the variability introduced by multiple authors, vocabularies and diseases provide a consumate methodology that should prove to be succesful.

Text classification has been used with success in the medical field to detect the service line that is treating a patient [3], the presence of an undiagnosed disease [8] as well as in previous work to assign diagnosis codes to medical records [5, 7]. In [3], support vector machines are shown to be a succesful classification tool for analyzing medical records and identifying the service line that treated the patient. In [5], the authors create an array of the most probable diagnosis codes, given the text, that could be accessed by the user to aid in the code assignment. This type of tool is already present in many of the electronic health records used in the medical industry. However, providing a list of the most likely candidates does not remove the subjectivity from the assignment process. In order for the MHS leadership to have reliable, objective codes across the spectrum of facilities, an objective automated process must be created. This thesis provides evidence that an automated text categorization technique, that can remove the subjectivity from current medical coding practice, is possible.

1.3 Overview of MHS

The Military Health System is a global medical network within the Department of Defense (DoD) that provides health care to all U.S. military personnel, veterans, and their dependents, worldwide. The MHS contains 59 hospitals, 364 health clinics and a $50 billion budget that provides care for approximately 9.6 million eligible beneficiaries. Navy Medicine, a subset of the MHS, is responsible for 3 million beneficiaries with a total budget of approximately $17 billion. Navy Medicine provides care to deployed service members at sea as well as in multiple field hospitals located around the world. Furthermore, Navy Medicine's most intense resource commitment is manifested in the care of dependents and retirees in traditional medical settings. As has been seen in the national healthcare debate, the delivery of medicine is fraught with inefficiencies and resource constraints that can mitigate a patient's ability to receive care. The leadership of Navy Medicine has developed a strategy to facilitate the way ahead outlined in the Navy Medicine Strategic Plan. The plan outlines objectives on the delivery of medical care in an efficient and effective manner.

1.4 Hospital Clinic Description

Naval Medical Center Portsmouth (NMCP) is a tertiary care medical facility that provides medical services to the Sailors and Marines stationed at Naval Base Norfolk, Naval Air Station Oceana and the surrounding military facilities. Naval Medical Center Portsmouth was selected as the contributing medical facility due to its volume of family practice encounters, its status as a teaching hospital, and the number of branch clinics that compose the medical delivery network. The medical network is comprised of ten branch clinics located throughout the greater Norfolk area. These clinics treat approximately 360,000 patients per year. Six months of data, which constituted 124,766 patient encounters, was collected from NMCP. These encounters are randomized and split into training, validation and test sets.

1.5 Thesis Outline

The second chapter details the data, the data systems and the method of procurement. Furthermore, a detailed description of the data fields included. The third chapter describes the Term Document Matrix Weighting schemes and the text classification methodology. The fourth chapter provides an analysis of the results of the classification analysis. The final chapter presents conclusions and recommendations for further research.

CHAPTER 2:
Data

The data for this study was procured from the Clinical Data Mart (CDM). Records from the family practice outpatient clinics of Naval Medical Center Portsmouth were extracted from the CDM. This location was chosen due to the high concentration of Navy personnel as well as the volume of medical care that is delivered on a daily basis. Data is delivered via Excel spreadsheet and is coerced into an extensible markup language (XML) format for easier entry into the R software package. Data was presented in two seperate categories. Initial patient demographic data consisted of approximately 124,000 records, where a record corresponds to a single patient encounter, and contains twelve separate data fields. In addition to the demographic data, textual data, matched to the corresponding patient encounter, was provided. File size requires that the data extraction from the CDM be separated into multiple transactions. Matching demographic data to its corresponding textual records is accomplished using the CDM Patient ID number and the Julian date of the encounter. Concatenating these data elements creates a single unique identifier for the complete record. Data coercion presents a significant challenge and will be discussed in detail in Chapter 3.

Data is partitioned into a training and test set. Eighty percent of the records are assigned to the training set and twenty percent to the test set. Partitioning is accomplished by the generation of a random number, distributed uniformly, between 0 and 1 where values less than or equal to 0.8 cause the record to be assigned to the train set and values greater than 0.8 cause assignment to the test set. The training set is partitioned a second time, using the methodology described above, in order to create a validation set. Thus, the training, validation and test sets are respectively 64%, 16%, and 20% of the original 124,766 records. The training set is used to build SVM clasification models. The validation set is used as a pre-test set in order to tune the parameters for each of the models built using the training set. The validation set provides the means to select the best candidate model. Classification on the test set provides the final results for this analysis.

In this chapter, we describe, in more detail, the source of the patient records and the CDM. We also describe the data fields and provide preliminary summary statistics. The reader should note that, though we choose to focus our classification methodologies on the textual data, summary statistics, demonstrating the demographics of the patient population, are provided for context.

Although the free-text physician notes, the focus of this thesis, form the basis of any diagnosis classification methodology, the intent of this thesis is to provide a foundation for further research. Therefore, we deem it prudent to also describe data elements that will be essential for enhanced classification research.

2.1 Data Sources

2.1.1 AHLTA

Armed Forces Health Longitudinal Technology Application (AHLTA) is the second generation of electronic medical records used by the MHS. AHLTA is a graphical user interface based system that builds upon the framework of the Composite Health Care System (CHCS). The intent of AHLTA is to provide a medium for physicians to record their patient encounters with as little free-text entry as necessary. Features such as automated documentation and checkboxes for symptoms are intended to enhance clinician efficiency while maintaining a high level of documentation accuracy. Efficiencies are laudible; however, selecting the final diagnosis code remains the responsibility of the physician and physicians are not coders. Physicians' lack of expertise, and motivation, provides the basis of all coding issues within the MHS. AHLTA provides workflow efficiencies to the physician; however, most physicians still provide a detailed description of the encounter in the free-text record.

2.1.2 CHCS

Physicians chart and document their work in the MHS utilizing an electronic medical record known as the CHCS. CHCS is an automated medical information system that provides documentation and support to clinicians through out the MHS. It is the backbone of the medical information system for the DoD, with AHLTA being the primary interface, and provides access to patient information at DoD hospitals and clinics around the world. Much of the data collected is in free-text format that can be used as reference by individual providers to learn the medical history of a patient. Histories are beneficial for single providers; however, the information contained in histories cannot be easily aggregated to provide insight for MHS leadership into the medical history of the beneficiary population.

2.1.3 Clinical Data Mart

The CDM was queried for a six month period, January 2010 to June 2010, producing 124,766 records for training and testing. The CDM is the end-user database that aggregates the raw medical information that is compiled in the two front-line systems, AHLTA and CHCS. The CDM

contains approximately 5.2 terabytes of medical data, ranging from free-text orders and notes to numerical data fields, that contain volume of medications, demographic data, and physician data.

2.2 Data Fields

Several types of information are extracted from the CDM for each record to ensure that multiple classification techniques could be attempted in the future. For example, demographic data, such as patient sex and age, are included in addition to the text field, containing physicians' notes, which is used for semantic-based classification. Again, the intent of this thesis is to provide a solid foundation for more enhanced classification techniques. We believe that demographic information and patient histories can be used to augment the SVM semantic-based classification techniques of this thesis. The patient information provided for this thesis contains the following fields for each patient encounter:

- Military Treatment Facility (MTF) Defense Medical Inforation System (DMIS) ID
- CDM Patient ID
- Family Member Prefix (FMP)
- Sex
- Patient Age
- Appointment Date
- Appointment Status
- Appointment Type
- Medical Expense and Reporting System (MEPRS) Code
- DMIS ID
- Clinic Name
- ICD9 Code
- Physician Note

The Physician Note is used to define the independent variables for the SVM classification models through the use of text processesing software. The ICD9 codes will comprise the dependent variables in the analysis. The remaining data fields are provided for future endeavors.

2.2.1 MTF DMIS ID

The Defense Medical Information System Identification number (DMIS ID) is a numeric identifier assigned to each of the medical facilities within DoD. This number is used to track patient enrollments as well as workload at each of the medical facilities around the world. In total there are over 9,000 DMIS ID's. Many of these are for field hospitals, battalion aid stations, as well as disease labs, and veterinary clinics in addition to hospitals and primary care clinics.

2.2.2 CDM Patient ID

The Clinical Data Mart Patient ID (CDM ID) is a unique identifier assigned to each of the patients in the MHS. This number can be used to track the care of a patient, longitudinally, as he or she progresses through the MHS continuum. The CDM ID allows for a non-patient-identifying field to be used instead of a Social Security Number. This study consists of 122,474 individual patients.

The CDM ID played a key role in matching patient demographic data to the corresponding physician note. In addition, the CDM ID provides the necessary granularity in the data to analyze longitudinal relationships among codes.

2.2.3 FMP

The Family Member Prefix (FMP) is a numerical representation of the patient's relationship to the active duty sponsor. The numerical representation of the patient's relationship allows for easy access for analysts to sort and filter for certain population types within studies. A summary of the FMP relationships is shown in Table 2.1.

2.2.4 Sex

The Sex field shows each the gender for each patient record. Gender is denoted by M for male and F for Female. A single record with an "unknown" categorization was present in the data. The record was removed. The distribution of Sex is shown in Table 2.2. It should be noted that Table 2.2 shows the frequency of gender by record while Table 2.1 shows the frequency of patient FMP by indvidual patient. The change in perspective results in the different frequency totals between Table 2.1 and Table 2.2.

FMP	Beneficiary Status	Frequency
01 - 19	Children	36,202
20	Sponsor	42,175
30 - 39	Spouse	43,977
40	Mother or Stepmother	54
45	Father or StepFather	2
50	Mother-in-Law	12
55	Father-in-Law	0
60 - 69	Other Authorized Dependent	37
90 - 95	Beneficiary Authorized by Statute	0
98	Civilian Emergencies	3
99	All Others, Not Elsewhere Classified	12
Total		122,474

Table 2.1: Family Member Prefix. Note that the frequency is a description of the number of individuals in the data, not necessarily the number of records, as has been described in other summary statistics.

Symbol	**Sex**	**Test**	**Train**	**Total**
M	Male	11,374	64,682	76,056
F	Female	7,398	41,311	48,709
U	Unknown	0	1	1
Total				124,766

Table 2.2: Distribution of Data by Sex. Note that the total describes the number of Male and Female records in the data set.

2.2.5 Patient Age

The Patient Age is the age of the patient, in years, at the time of the encounter with the physician. Figure 2.1 provides the distribution of beneficiaries by age. It appears that certain age groups require more care than others. This is to be expected. Infants require frequent well baby checkups, hence the high frequency at the age of 1. Again we see an increase in physician visits from age 24 to 35. This age group makes up the largest population in the United States Military and, therefore, provide the largest resource demand the MHS.

2.2.6 Appointment Date

The Appointment Date is the date on which the encounter with the physician occured. Data was only extracted for patients seen between January 1, 2010 and June 30, 2010.

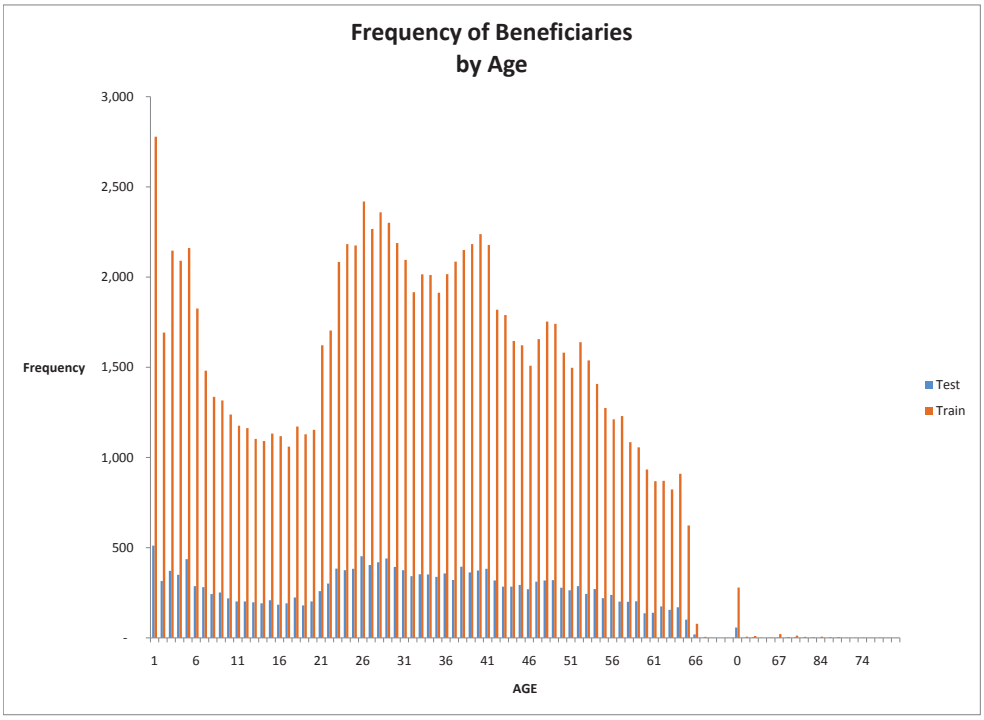


Figure 2.1: A histogram of the frequency of beneficiaries by age contained in the data. The orange lines represent the distribution by age for the training data set and the blue lines represent the distribution by age for the test data set.

2.2.7 Appointment Status

The Appointment Status is a marker that identifies the status of an appointment. The statuses include Kept, No-Show, Cancel, and Telephone consults. Each status provides information as to how the patient treated the time slot alloted to him or her.

2.2.8 Appointment Type

Appointment Types are used by the MTF to ensure that a physician has an array of different types of availability and to ensure access by all beneficiaries to the system. The appointment types are shown in Table 2.3.

2.2.9 MEPRS Code

Medical Expense and Reporting System (MEPRS) codes are four-level codes used to represent work centers within each MTF across the MHS. Initially these codes were utilized for cost ac-

Type	Description	Test	Train	Total
Acute	Patients that must be seen today	281	1,501	1,782
Follow Up	Patients that have been seen and require a checkup	6,105	35,236	41,341
Open Access	Any patient who needs care	6,660	37,242	43,902
Routine	Patients who have illnesses that are not urgent	81	428	509
Wellness	Physicals, yearly exams	3,208	17,661	20,869
Group	Appointments in a group setting	19	68	87
Inital Spec. Care	First visit with a Specialist	2	33	35
Procedure	A procedure will take place	94	520	614
Tel. Consult	Patient recieves diagnosis via Telephone	2,322	13,305	15,627

Table 2.3: Distribution of Data by Appointment Type

counting purposes in order to ensure the tracking of costs in an MTF. However, in recent years MEPRS codes have also been used to track the workload generated in each work center. The MEPRS Code breaks down into four seperate characters. The first character, or level, describes the overarching purpose of the workspace or the Functional Category. An **A** represents an in-patient space, a **B** represents an outpatient space, a **C** represents Dental, and so on. The next character of the code, or second level, represents the type of work being done or the Summary Account. An **A** represents medical space, a **B** represents surgical, etc.... The preponderance of data are represented by the Functional Category **B** and the Summary Account of **H**. **BH** describes primary care clinics and are extended by a third-level code. The third-level is the work center description and the fourth level describes the type of work done in each space. For example, BACA represents an outpatient clinic, classified as Medical work with a specialty of Cardiology with a standard visit. The creation and management of these codes allow for hospitals within the MHS to compare workload as well as to provide meaningful reports of productivity to upper management at many levels. In this study they may be used as a demographic seperator. Table 2.4 shows the distribution of records for each of the MEPRS codes.

2.2.10 Clinic Name

A human readable interpretation of the MEPRS code that describes the location where the work load occured. The clinic names can be seen in Table 2.4 as well as the relationship between the MEPRS codes and the clinic names. Each **BH** clinic is denoted as a primary care clinic; however, the third-level MEPRS code, the work center, is different for each. This difference manifests itself in the different locations for each of the clinics.

MEPRS Code	Description	Test	Train	Total
BGAA	Family Practice Clinic	148	882	1030
BHA2	Deployment Health Clinics	189	1100	1,289
BHAA	Preventive Medicine Clinic	4	23	27
BHAJ	Primary Care Clinic - Boone	5,209	29,305	34,514
BHAK	Primary Care Clinic - Oceana	3,666	20,487	24,153
BHAM	Primary Care Clinic - Northwest		2	2
BHAO	Primary Care Clinic - MSC DN	1	10	11
BHAR	Primary Care Clinic - MSC NNY	1	3	4
BHAS	Primary Care Clinic - Sewells	2,241	13,011	15,252
BHAT	Primary Care Clinic - Chesapeake	2,956	16,823	19,779
BHAV	Primary Care Clinic - VA Beach	4,356	24,347	28,703
BHAY	Primary Care Clinic - BMC Yorktown	1	1	2

Table 2.4: Distribution of Data by MEPRS Code

2.2.11 ICD9 Code

International Classification of Diseases, ninth revision (ICD9), codes are used to code and classify morbidity data from the inpatient and outpatient records, physician offices, and most National Center for Health Statistics surveys. They are the official system of assigning codes to diagnoses and procedures associated with hospital and clinic utilization in the United States. Figure 2.2 provides insight into the distribution of codes.

ICD9 codes present a straightforward nominal numeric structure. The code is built hierarchically with the primary disease represented by the whole number, the code root, and additional specific information can be found in the decimal place, an extender. For example, 403 is Essential Hypertension and can be expanded upon by adding a 0 to describe the code as malignant. Thus 403.0 is malignant essential hypertension and 403.1 is benign essential hypertension. The hierarchy of information in ICD9 codes can be exploited to simplify classification problems by removing the extenders and dealing only with the code roots, thus reducing the total number of categories to be classified.

2.2.12 Physician Note

Physician notes are a compilation of multiple fields from the AHLTA patient record. The note is a text description of the patient encounter, physician's opinions, and diagnosis of the patient ailment. Furthemore, it includes patient history and the chief complaint for each encounter. The patient note contains an array of information, much of which is encoded in the cryptic shorthand

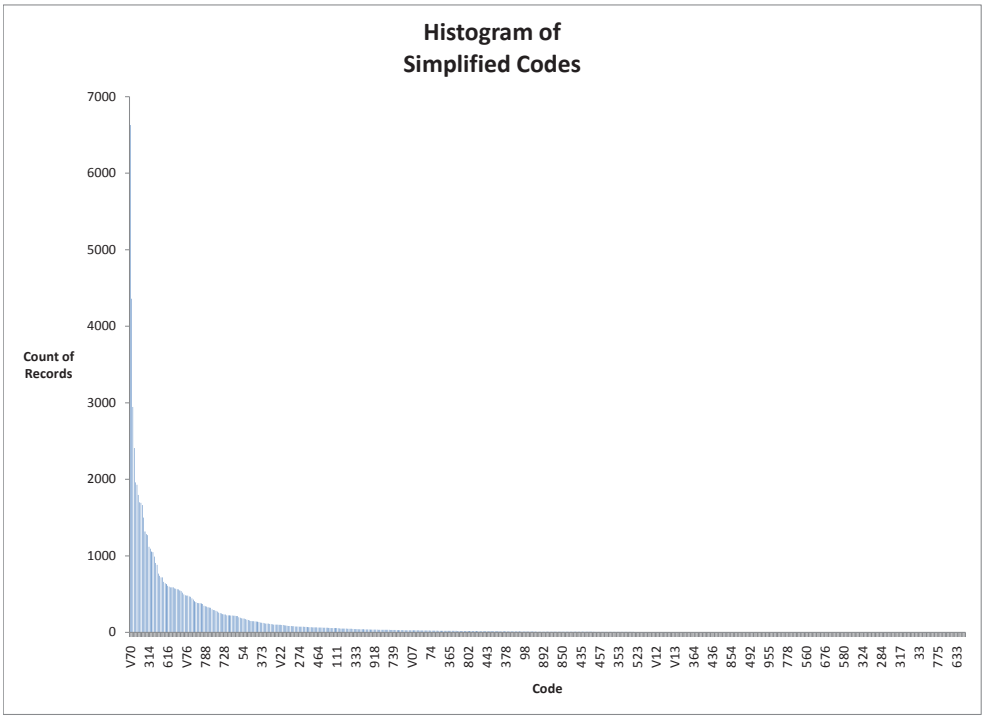


Figure 2.2: A histogram of the ICD9 codes presented in the training data set. We show this histrogram to provide context to the distribution of codes in the data set. Due to the large number of infrequent codes in the data, as well as the memory requirements for data processing, we choose to focuse on the 20 codes. These codes can be viewed in Figure 3.5

of a physician. Examples of notes include:

> pt is a 30 y/o whi c/o 25lb weight gain in 5 weeks. has a h/o TAH ovaries remain in Oct for fibroids and DUB. has a h/o primary Hypothroidism on Synthroid, dose stable for 4 years. she quit smoking with chantix recently and is worried ince she ahs been more physically active and keeps gaining. she feels tired, acne has flared as well. weight gain

and

> pt here for CPE he states has appt with PT for elbo and knee pain and has appt for sleep study. Oter wise djng well without c/o. He has tried nicotene patch and

gum and has not worked and would like to stry something else. no psych hx

The two examples illustrate a number of difficulties within the data: shorthand, jargon, misspellings and overall short descriptions. Unfortunately, these records are not the most challenging physician notes in the data set to classify. Due to the automated nature of AHLTA a number of records contain little to no discriminating information. For example:

Allergy information in Autocite area was reviewed and verified with patient. Curent medications reviewed and reconciled. :Reviewed

Records containing little discriminating information are not likely to be classified correctly; however, they represent a key area of interest for this study.

CHAPTER 3:
Model Formulation

In this thesis, free-text disease classification is accomplished using the support vector machine construct. In order to use this supervised learning methodology a series of steps, described in the following section detailing the process and assumptions that are incorporated into the classification model. The steps begin with the extraction of the data from the CDM, followed by the delivery and conversion from multiple Excel files into a single XML source. Subsequently, the XML source is processed by R using the software package tm [9], undergoes feature selection and culminates with the fitting of SVM models. We close the chapter with a discussion of the performance metrics employed to measure the effectiveness of each classification scheme.

Medical records are delivered in an Excel format, as is mentioned in Chapter 2. Each medical record consists of multiple rows in Excel requiring aggregation into a single record. Visual Basic for Application (VBA) scripts were written to perform the aggregation task programmatically. Once aggregated, each row represents a patient encounter. Upon completion, the data is formatted into a XML document. A custom XML schema was created and applied to the aggregated data in order to produce the final XML document. With the aggregated data formatted, the data was exported into XML files. The XML files were read into PYTHON to remove a number of ASCII characters that were not recognised by R. These characters included: end of transmission, bell, device control 1. An R add-on package, tm[9], is used in conjuction with a custom XML parser, supported by the XML package[10], to convert the XML documents into plain text documents, held within a corpus in memory. The computational requirements for processing the medical records and storing the corpus are extensive. As a frame of reference, the training set requires 4GB of RAM to hold its corpus in memory. Additional work, such as fitting support vector machine models, requires supplementary computational assets.

Text processing methodologies, resident to the tm package, are used to strip extra white space, convert letters to lowercase, remove suffixes, and remove stop words from each document. Stop words are those terms used in sentences, such as “a,” “the,” “that,” etc. . . , that are used so often they provide no discriminatory power for a classification model. Text processing methodologies are the first step in feature selection, a technique described later in this chapter. The final data processing step is to convert plain text documents, still written in prose, to a mathematically friendly data structure. This thesis uses the Term Document Matrix (TDM) for such a purpose.

The TDM is a construct resident to the tm package of R and provides great flexibility for regular textual analytics. TDM's store data in a simple triplet matrix representation. This representation stores each value as a triplet where the first value represents the row number, the second value represents the column number and the third value represents the value contained within the matrix. Previous analytical work using the tm package was based on smaller data sets that could utilize the simple triplet matrix framework and the RAM requirments of holding the TDM matrix was not prohibitive.[11] Due to the size of the training corpus used in this thesis, conversion of TDM data structures to compressed sparse row was necessary. Conversion of the TDM allows the matrix to be used by the SVM software [12]; however, the new data representation reduces the metadata that is carried along with the compressed version of the TDM. Therefore, a series of metadata functions were written and applied to the corpus to create vectors for each of the demographic categories. These vectors can be applied in order to reconstitute the wealth of information available in the original data structure.

3.1 Term Document Matrix

The TDM is the key data structure when developing a text classifier. Its construction involves analyzing the corpus of medical records and transforming each document into a vector of term counts or term frequencies (tf). An example of such a document vector is

$$\begin{bmatrix} \#\{Reviewed\} \\ \#\{Xray\} \\ \#\{Patient\} \\ \vdots \\ \#\{Infection\} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 3 \\ \vdots \\ 1 \end{bmatrix}$$

in which the term "Reviewed" appears once, "Xray" does not appear, and so on. These vectors are used to construct the TDM whose elements are functions of the term frequencies. Each row of the TDM represents a term from the corpus and each column represents a document. Specifically, the elements, x_{ij}, of a TDM are associated with the occurence of term i, in document j. That is, x_{ij} might be taken to be the frequency of the i^{th} word in the j^{th} document. As one might assume, the TDM can be very sparse and in many instances individual terms are used only once in the corpus. In addition, document authors have different vocabularies and patterns of term usage. The value of each x_{ij} can also be calculated using multiple weight schemas which apply

weights to the term frequencies, also known as feature selection techniques. Feature selection is the process of identifying terms that are the most useful in determining the differences between classes of documents and includes, not only term weighting, but term selection.

3.2 Feature Selection

Feature selection increases the efficiency and the effectiveness of the classifier by enhancing the similarities and the differences between like and unlike document vectors, respectively, by normalizing the term frequencies and reducing the feature space. Studies have shown that large reduction in feature space has little negative effect on the accuracy of a classifier. [6] Document frequency techniques were shown to be an effective selection criteria when dealing with large data sets. We will use two techniques, term selection and term weighting, for feature selection in this thesis. These techniques have been used succesfully in previous medical text classification. [3]

3.2.1 Term Selection

Term selection is an important step in developing the feature space for the text classifier. Terms that appear in extremes, either frequently or infrequently, are less likely to enhance the ability of the model to discriminate between classes. Therefore, sets of terms, falling into the extremes, should be eliminated to increase the efficiency of the classification algorithm. Zipf's law [13] provides a framework for the distribution of term frequencies in a corpus. Furthermore, Zipf's law provides a structure for the removal of very infrequent terms as well as terms that are deemed to occur too frequently. Figure 3.1 shows the distribution of term counts for the training set of data. In Figure 3.1, terms are first ranked according to their total term frequency over all documents tf_i $i = 1, \ldots, I$ where I is the number of terms in the corpus, the y-axis is $\log(tf_i)$ and the x-axis is the log rank of each term. The logs of the frequency and ranks are used due to the extreme values in the data set.

The training set has approximately 81,000 different terms. Many of these terms are used often. For example, the word "reviewed" is used 293,126 times. That averages 3.5 uses per document in the corpus. We believe the prevalance of the term "reviewed" results from those records that contain auto-populated segments, or that require some sort of acknowledgement of review. While it is helpful to have an overlap of terms, it is unlikely that the word "reviewed" will provide any discrimination between classes. In addition, Figure 3.2 shows 45,981 terms that are used once. After inspection we find that most of these terms are errors in typing that include miss-spellings and forgotten spaces. These errors are removed from the training data set as it

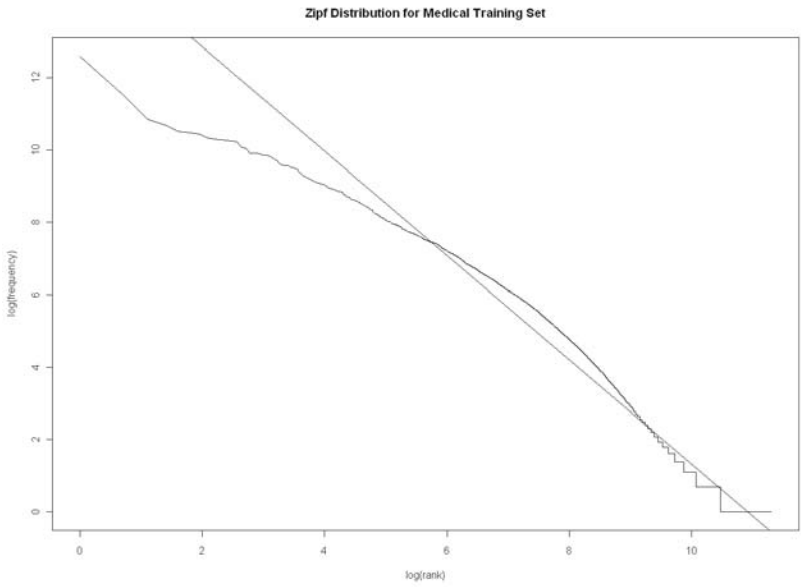


Figure 3.1: A Zipf distribution of the frequncy of terms used in the training set where terms are ranked according to frequency. The distribution shows a large number of terms that are used infrequently

is unlikely that they will provide any discrimating evidence to the classification model yet to be trained. Previous works [3, 4] use document term frequency cutoff of three which reduces their feature spaces by 60%. If that criteria is used in this analysis, 76% of the terms would be removed. Therefore, a term frequency cutoff of two is used, reducing the feature space by 70% or 57,670 terms. In addition to a low-end cutoff, a ceiling usage will be sought. A cutoff of 84,919 was selected as it represents a word that is used at least once per document in the training set, where the training set contains 84,919 documents. The ceiling requirement only removes two terms from the dictionary; the previously mentioned "reviewed" as well as the word "patient".

3.2.2 Term Weighting

While term frequency provides gains in enhancing the computational efficiency of the classifier, term-weighting provides a means to enhance the information gain in each terms usage. Term-weighting schemes enhance an algorithm's ability to retrieve documents by emphasizing the characteristics unique to the document, and minimizing the characteristics that are similar to those in other documents. [14] The same ideas have been adopted by text classifiers. Three weight schemas, term frequency, term frequency-inverse document frequency, and normalized term frequency-inverse document frequency come standard in the tm package of R. [9] These three weighting schemas have also proven to be satisfactory weight schemas in other research.

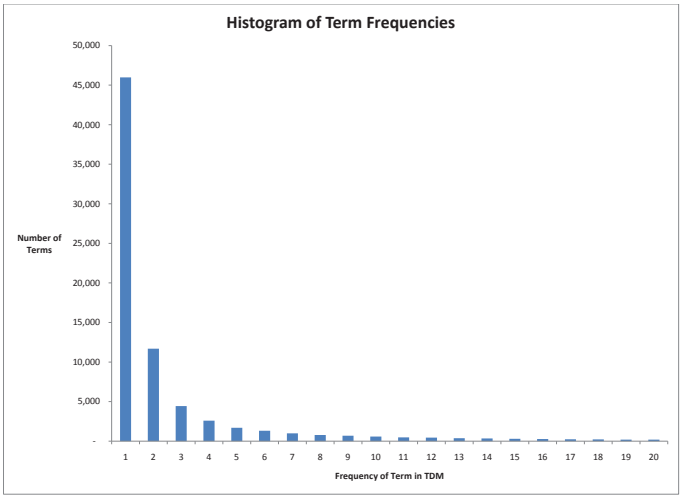


Figure 3.2: A historgram of term freqeuncies illustrates the number of terms used infrequently. 45,981 terms are used once and 11,689 terms are used twice. Therefore, terms used twice or less represent 70% of the terms and terms used three times or less represent 76% of the terms

[3, 4]

Salton and Buckley [14] propose a construct for enhancing term importance in a TDM that involves three different perspectives on the relationship of a term with its class.[14] The relationship can be described using three components: local, global, and normalization. These three components are used in the three weighting schema. The local component describes the importance of a term within a single document, where importance is measured by frequency. The local component is term frequency, tf_{ij}, of each word in each document. The global component looks across the corpus of documents and describes how important the term is across all of the documents. Finally, the normalization component seeks to adjust for documents that use more terms than others.

Term Frequency Schema

Let I be the number of terms in the corpus and J be the number of documents. The term frequency schema uses the number of times a term, t_i, is used in a document, d_j.

$$tf_{ij} = \#\{t_i \in d_j\} \quad \textit{for i=1...I,for j =1...J} \tag{3.1}$$

The term frequency provides the integer values for each term's usage. It is the rawest form of term weighting and does nothing to adjust for a term's usage outside of a given document. The term frequency schema is based soley on the local component.

Term Frequency-Inverse Document Frequency

While the local factor describes the importance of a word within a single document, a global factor expresses the relative importance of a word across all of the documents in the TDM. As such, the global factor has the ability to discern if a word is used too much or too little in the dataset to be useful. Let N be the total frequency of all terms in the corpus i.e.

$$N = \sum_{i=1}^{I} \sum_{j=1}^{J} tf_{ij} \tag{3.2}$$

The inverse document frequency, idf_i, is defined for each document $i = 1 \dots I$ to be

$$idf_i = \log_2 \left(\frac{N}{\sum_{j=1}^{J} (tf_{ij})} \right) \tag{3.3}$$

The inverse document frequency is a global factor because it varies inversely with the number of documents to which a term is assigned in a collection of J documents.[14] It serves as an adjustment to the term frequency factor in an effort to reduce the impact of terms that are popular throughout the corpus. In the term frequency-inverse document frequency schema, the elements of the TDM are taken to be

$$tfidf_{ij} = tf_{ij} \cdot idf_i \quad \textit{for i=1, \dots, I, j = 1, \dots, J} \tag{3.4}$$

Normalized Frequency-Inverse Document Frequency

A normalization factor is utilized to adjust for the variability in the number of terms used in each document. For example, longer documents tend to use a larger variety of terms. Functionally, a document which contains more terms is more likely to be considered "important" when creating

the classification scheme simply because it has more terms in common with other documents. The normalization component, as the name suggests, provides a system to normalize the effect each document will have on the classification algorithm. The normalization is computed by dividing the term frequency by the total number of times each term appears in the document. The entries in the TDM for this schema are

$$TfIdfN_{ij} = \frac{tf_{ij}}{\sum_{i=1}^{I} tf_{ij}} \cdot idf_i \quad \textit{for i = 1, \ldots, I, j = 1, \ldots, J} \tag{3.5}$$

The components in equations (3.2), (3.4) and (3.5) are used to build TDM's. These approaches were originally developed based on the empirical observations that: (1) the more times a word appers in a document, the more relevant it is to the subject of the document, and (2) the more times a word occurs throughout all the documents in the collection, the more poorly it discriminates between documents.[14]

3.3 Text Classification Model

A number of classification techniques have been developed in an effort to classify text documents in many settings. This thesis will focus on a vector space approach using support vector machines.

3.3.1 Support Vector Machines

Support Vector Machines (SVM) are a supervised statistical learning technique introduced by Corinna Cortes and Vladimir Vapnik to classify observations by mapping their "input vectors into a high dimensional space, Z, through some non-linear mapping." [15] The SVM seeks to find a hyperplane that maximizes the margin between two classes in a high-dimensional space. This hyperplane is known as the optimal hyperplane. For example, consider the two classes represented by orange plus signs and blue hyphens in Figure 3.3. The goal is to find the linear function of the two variables x_1 and x_2 which yields the widest separation, or largest margin, between the two classes. In Figure 3.3, the optimal seperating hyperplane is in fact a line. The margin is indicated by the dashed lines parallel to the optimal seperating hyperplane. The points on the margins are called the support vectors.

In Figure 3.3, the two classes are separable in two dimensions. Classes are not always separable

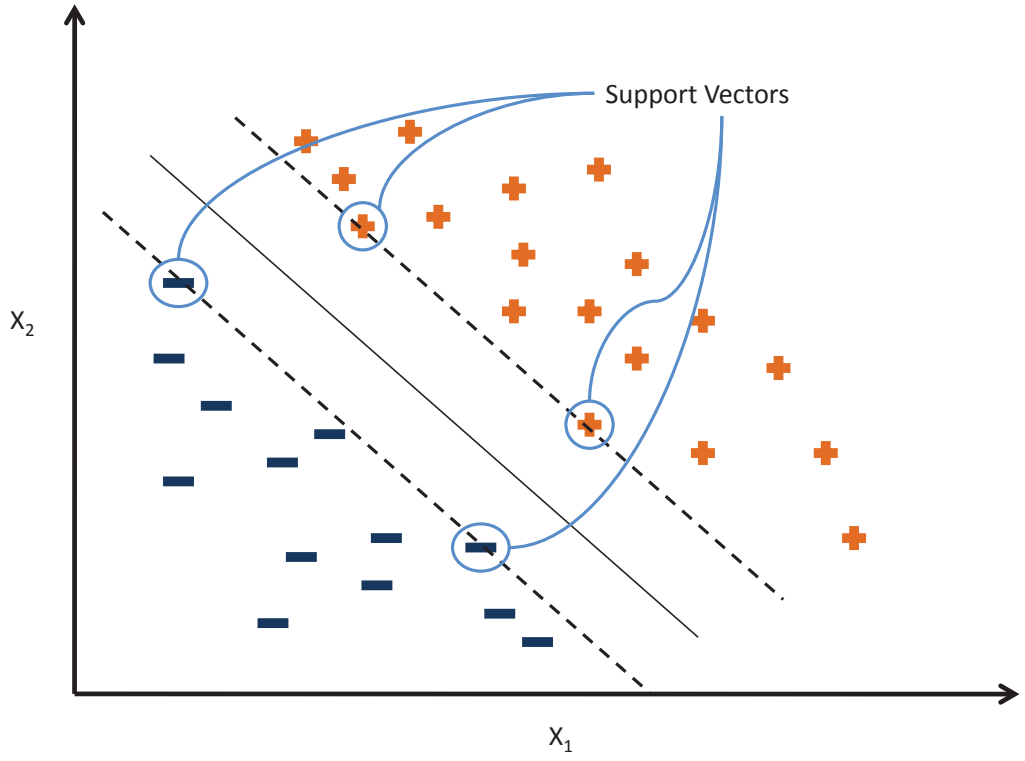


Figure 3.3: Example of a linearly separable problem in 2 dimensions. Support vectors define the margin of largest separation between the two classes.

in their original space. By mapping them to a sufficiently high-dimensional space, classes can be made seperable. SVM classification based on seperable classes without any crossover produces what is known as a hard margin classifier. Hard margin classifiers, based on sufficiently high dimensional spaces are usually poor classifiers. [2]

Classifying lower-dimensional spaces in which the classes are not seperable requires a slightly different formulation. For clarity of purpose, we diverge from the common notation used to describe SVMs, letting J represent the number of documents, or observations, and I represents the terms or the dimension of the "input vectors". The SVM accomplishes this by introducing slack variables $\xi_j \;\; j=1,\ldots, J$. The slack variable allows observations to be on the "wrong side" of the seperating hyperplane with a penalty against the objective function. Figure 3.4 shows an example in two dimensions where the data is not linearly seperable.

Specifically, in I-dimensional space we wish to find the hyperplane $\beta_0 + \beta_1 x_1 + \ldots + \beta_I x_I$

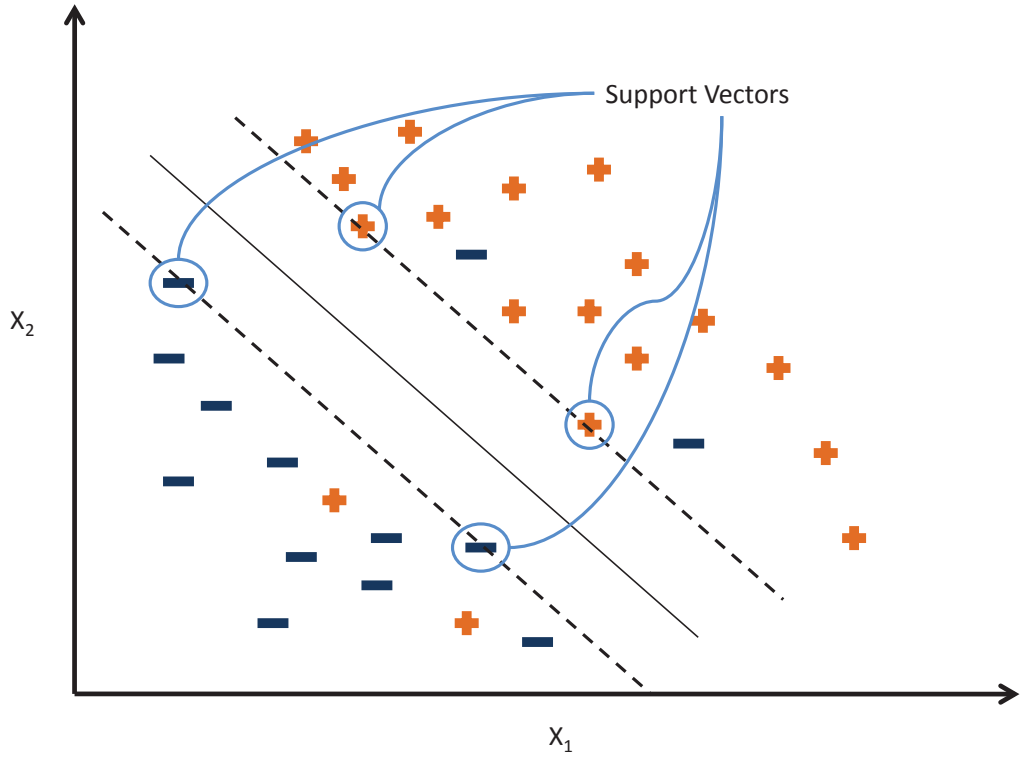


Figure 3.4: Example of a Non-linearly separable problem in 2 dimensions. Support vectors define the margin of largest separation between the two classes.

where $\sum_{i=1}^{I} \beta_i^2 = 1$ which maximizes the margin M subject to constraints

$$y_j(x_j^t\beta + \beta_0) \geq M - \xi_j \quad for\, j = 1,\ldots,J \tag{3.6}$$

$$\xi_j \geq 0 \quad for\, j = 1,\ldots,J \tag{3.7}$$

and

$$\sum_{j=1}^{J} \xi_j \leq constant \tag{3.8}$$

where for $j = 1, \ldots, J$ $y_j = +1$ or -1 depending on the class of the j^{th} observation $x_j^t = (x_{1j}, \ldots, x_{Ij})$ is the input vector for the j^{th} observation and $\beta^t = (\beta_i, \ldots, \beta_I)$, the vector of coefficeints for the seperating hyperplane. Equivalently, the problem of finding the optimal seperating hyperplane is a quadratic minimization problem with linear constraints:

$$\min_{\beta_0, \beta} \frac{1}{2}\|\beta\|^2 + C\sum_{j=1}^{J} \xi_j \tag{3.9}$$

Subject to:

$$y_j[x_j^T\beta + \beta_0] \geq 1 - \xi_j \quad \textit{for j=1,...,J} \tag{3.10}$$

$$\xi_j \geq 0 \quad \textit{for j=1,...,J} \tag{3.11}$$

where C is the cost, and $C\sum_j^J \xi_j$ is the soft margin function. The cost allows the user to vary the penalty by which a point on the "wrong" side of the hyperplane will affect the value of the objective function. Small values of C increase the allowance for misclassification of the training data (or training error). A large value of C leads to behavior more similar to a hard margin SVM. The result is very little training error, but classification models that perform poorly on test data. While most articles concerning SVM text categorization acknowledge that the cost parameter must be tuned, no guidance could be found for initial starting values. Furthermore, [2] argues that when the number of terms is much larger than the number of documents, the advantages of using soft margins is decreased. While in this thesis I is extremely large, it is believed that the diagnosis codes will not be strictly linearly seperable and, therefore, we initially model cost values $C \quad \in \{2^{-5}, 2^{-3}, 2^{-1}, 2, 2^2, 2^3\}$.

Classifying diagnosis codes based on patient record documents requires a seperate SVM to be fit for each of the chosen diagnosis codes included in this study. Each SVM classifier will assign to each record a probability that the record belongs to the code in question. These probabilities will be aggregated, and the most likely code, the one with the highest probability, will be selected for each record. Adding another level of refinement allows us to artificially increase the precision, a measure to be discussed in the next section, and increase the overall performance of the classifier. The additional level of refinement will be a confidence cutoff that will ensure that the probability produced from the SVM is above a certain threshhold in order for a code to be considered as the predicted code. For example, if the confidence cutoff is 0.8 and the largest probability, say for code V70, is 0.7, then the record will be labeled as "other". We vary the cost variable as well as the confidence cutoff to find the best classifier for each code.

In order to simplify the classification problem, the complexity of the classificaiton is reduced by focusing on the roots, the whole numbers of the codes, and eliminating the extenders, the decimal points, reducing the total number of codes from 2,265 to 624. Furthermore, for ease of purpose, and due to time constraints, only the 20 most numerous codes are included. Initial tests show that each code's SVM model could take up to 8 hours to fit, requiring up to 4 months to fit the 360 models required for this thesis. Figure 3.5 provides the frequencies for the 20 codes tested in this thesis. In addition, Table 3.1 presents a description of each code, the frequencies amongst the test and training sets, as well as the overall frequency of each code.

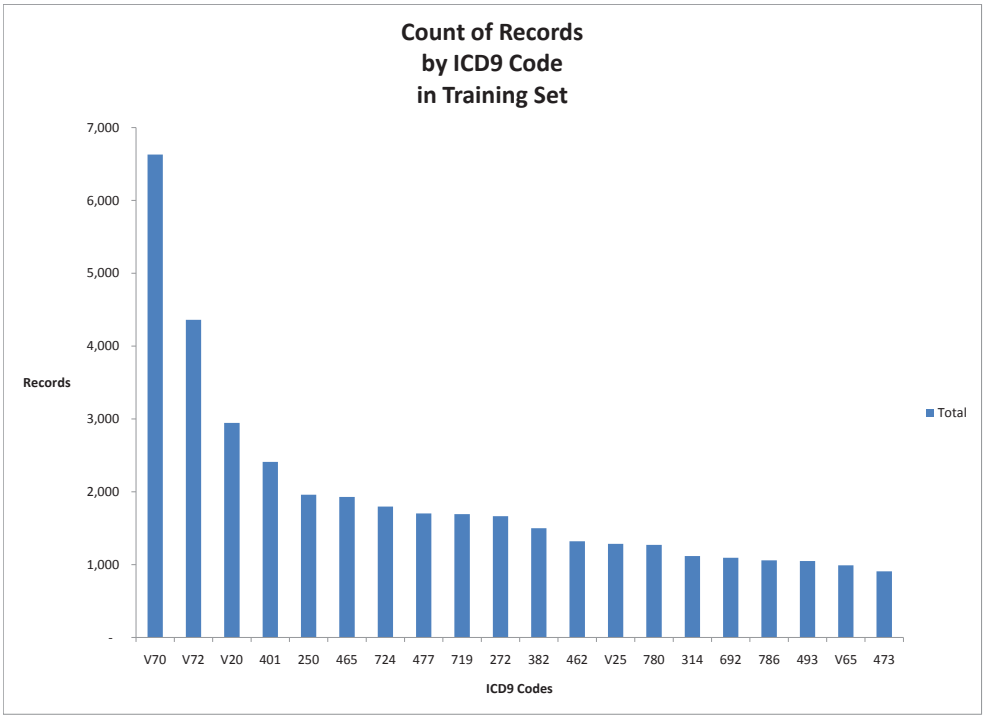


Figure 3.5: A histogram of the top 20 ICD9 codes in the training set. V70, General Medical Examination, is shown to be the most common with more than 6,500 records in the training set. Chronic Sinusitis, Code 473, is in the 20th spot with just under 1,000 records. The codes selected represent a variety of different diagnosis from Hyperkinetic Syndrom of Childhood to Contact Dermatitis to Hypertension to Diabetes. The codes and their descriptions can be viewed in Table 3.1. The spectrum of codes present should provide some indication of the ability of the SVM to classify across many specialties.

ICD9 Code	Diagnosis Description	Test	Train	Total
V70	General Medical Examination	1,549	9,575	11,252
V72	Special Investigations	982	5,525	6,507
V20	Well Child Visit	633	3,686	4,319
401	Hypertension	571	2,991	3,562
465	Upper Respiratory Infection	431	2,447	2878
250	Diabetes Mellitus	398	2,413	2,811
724	Disorders of the Back	381	2,250	2,631
477	Allergic Rhinitis	390	2,153	2,543
719	Disorders of the Joint	370	2,140	2,510
272	Disorders of Lipoid Metabolism	390	2,118	2,508
382	Otitis Media	358	1,858	2,216
462	Acute Pharyngitis	292	1,626	1,918
V25	Contraceptive Management	292	1,593	1,885
780	General Symptoms	270	1,579	1,849
692	Contact Dermatitis	274	1,378	1,652
314	Hyperkinetic Syndrom Childhood	248	1,366	1,614
786	Symptoms Involving Respiratory System and other Chest Symptoms	227	1,329	1,556
493	Asthma	221	1,314	1,535
V65	Other Persons Seeking Consultation	217	1,259	1,476
473	Sinusitis	210	1,167	1,377
Total		8,704	49,767	58,699

Table 3.1: Presents the distribution of diagnosis codes across the different data sets. Furthermore, a description of each code is presented for future reference.

3.4 Measures of Success

In order to determine how successfully each classifier performs on the validation and test set, a standard set of four measures will be utilized. [3] These measures include accuracy, precision, recall and F-score. The first three measures in themselves are not enough to properly describe how well each classifer performs; the F-score is an overarching measure that evenly weights precision and recall. F-score is not unduly influenced by a large number of negative exmples, or records in the traning set which do not have the code corresponding to the classifier being assessed. For this thesis the F-score will serve as the overall measure of success because we seek to build an automated classifier. Building this classifier will require confidence that the predictions are accurate, precision, and that the classifier predicts enough of the codes correctly that are present in the population, recall, to be useful to the MHS. Therefore, F-score is a metric that provides an indsight into each of the requirements.

3.4.1 Accuracy

Accuracy is an intuitive measure for the performance for classification algorithms; however, it can be misleading. In cases where positive examples, i.e., records in the training set with the diagnosis code being classified, are scarce, the classifier gets a lot of credit for labeling negative examples correctly. Confusion matrices are constructed for each of the 20 diagnosis codes. A confusion matrix tabulates records by their observed diagnosis code ("+" if they have the particular diagnosis code, "-" if they do not) and by how the code's classification is predicted ("+" predicted with the code, "-" not). The elements of a confusion matrix are denoted by $f_{--}, f_{-+}, f_{+-}, f_{++}$ as depicted below:

	Observed = "+"	Observed = "-"
Prediction = "+"	f_{++}	f_{+-}
Prediction = "-"	f_{-+}	f_{--}

Accuracy is most commonly measured as the sum of the diagonal of the confusion matrix divided by the sum of the elements in the confusion matrix.

$$Accuracy = \frac{f_{++} + f_{--}}{f_{++} + f_{+-} + f_{-+} + f_{--}} \tag{3.12}$$

While accuracy can be misleading when few positive examples are available, precision and recall (below) are less sensitive to the total number of records classified and their focus in on how well the classifier performs when it labels a record as positive.

3.4.2 Precision

Precision, p, measures how often the classifier is correct when it labels a test record as positive. [3]

$$p = \frac{f_{++}}{f_{++} + f_{+-}} \tag{3.13}$$

3.4.3 Recall

Where precision measures how well the classifier performs within the realm of those records that were predicted to be positive, recall, r, assesses how well the classifer picks out true records. [3]

$$r = \frac{f_{++}}{f_{++} + f_{-+}} \tag{3.14}$$

3.4.4 F-score

Finally, F-score provides a summary statistic that balances the precision and the recall statistics of each classifier. In this thesis, the harmonic mean between of the precision and the recall will be utilized. [3]

$$F\text{-}score = \frac{2}{\frac{1}{p} + \frac{1}{r}} \tag{3.15}$$

3.4.5 Macro-Averaging

Macro-averaging corresponds to the standard way of computing an average. A performance measure (i.e. precision, recall, F-score, etc..) is computed separately for each classified value, F_i where F_i $i = 1, \ldots, 20$ are the F-score for the i^{th} classifier. The average is computed as the arithmetic mean of the performance measure over all classifiers. The F-score, being our main decision metric, would have a macro-average function of:

$$Macro(F\text{-}score) = \frac{1}{20} \sum_{i=1}^{20} F_i \tag{3.16}$$

CHAPTER 4:
Results and Analysis

Classification models are built for each of the twenty selected ICD9 codes using a range of tuning parameters. This chapter discusses the strengths and weaknesses of each of the proposed models as well as provide a discussion of the final model that was selected. The discussion of the models is broken down across the three weighting techniques described in Chapter 3 and focus on how the changes in tuning parameters affect the results.

The SVM models are evaluated using the validation set described in Chapter 2. Six models are constructed per code, totaling 120 models per weighting schema and 360 models for the thesis. Each model produces a probability describing how likely the record in question is to have the code for which the SVM was fitted. A probability is produced for each of the records in the validation set and for each of the SVM models. These probabilities are stored in a matrix where the columns represent the results for a codes specific model and the rows represent the document that the SVM is attempting to classify reulting in a matrix that is approximately 21,000, the number of records in the validation set, by 20, the number of codes for which we are testing. Once the 20 models are evaluated on the validation set, a series of processes determines the most likely code by finding the column for each document with the largest probability. We select the largest probability because it indicates the model that believes it has the "most right" prediction based upon the training set of data. While it would be simple enough to assign the code with the highest probability, a further layer of discretion is added. We choose to add another variable in the classification known as the confidence cutoff. The cutoff is a threshold that must be surpassed in order to classify a given document as any code. For example, if we set the cutoff to 0.8, then we are ensuring that the model is at least 80% confident that its assignment is correct. A larger confidence cutoff leads to higher precision values because it artificially increases how confident the SVM needs to be in order to assign a code to a record. Ultimately, the management of the precision and recall metrics allow a greater opportunity to build a successful classification model.

4.1 Initial Modeling

Literature review suggests using a series of cost variables, $C \in \{2^{-5}, 2^{-3}, 2^{-1}, 2, 2^2, 2^3\}$, and multiple weighting techniques would provide the best classification opportunity. [3, 4,

5, 6, 7] We begin the process with the understanding from previous work [2] that large data sets could classify well using hard margin classifiers, resulting from large cost variables, and using sophisticated weighting techniques. The term frequency-inverse document frequency and normalized term frequency-inverse document frequency weighting methodology have proven to be the most effective at classifying medical text. [3, 5, 7] Therefore, we begin our discussion with the more complex weighting schemas and progress to the simpler term frequency schema.

4.1.1 Term Frequency-Inverse Document Frequency

The first weighting schema, term frequency-inverse document frequency, presents a sophisticated feature selection technique that should enhance the similarities and differences between documents. The term frequency-inverse document frequency validation matrix is evaluated for six values of the cost parameter. A precision, recall and F-score metric is reported for each scenario. The first confidence cutoff is 0.8, indicating that the model must be at least 80% sure that the predicted code is correct before the predicted code could be assigned to the record. It is hoped that the size of the data set allows this high confidence level to prove useful. As a reminder, the F-score is the harmonic mean between precision and recall. Figure 4.1 shows the initial F-score values for the first weighting schema as a function of the cost parameter C.

These results require further examination and suggest a need for further refinement in our classification methodology. In order to create a plan for enhancing the classifers F-score, we consider the F-score's initial pieces, precision and recall. Remember, precision is the probability that the predicted classification is correct, given that it classified the record as positive. Recall is the fraction of correct classifications out of the population of positive examples.

The plot of precision against the cost parameter in Figure 4.2(upper panel), provides encouragement that an autoclassifier may in fact be possible. A handful of codes (V70, V72, 724, 473, *etc.*) have precision values above the 0.8 requirement and many have precision values at or near 1, indicating that every code that is predicted correctly. While this provides some encouragement, it also means that the recall values must be very small. The analysis continues with an inspection of the recall values.

The recall values, also presented in Figure 4.2(lower panel), indicate a very poor classifier. Many of the recall values fall to 0 values greater than 2^{-5}. Furthermore, for those codes whose recall values that don't fall to 0, their are still so low that they significantly drag down the overall F-score.

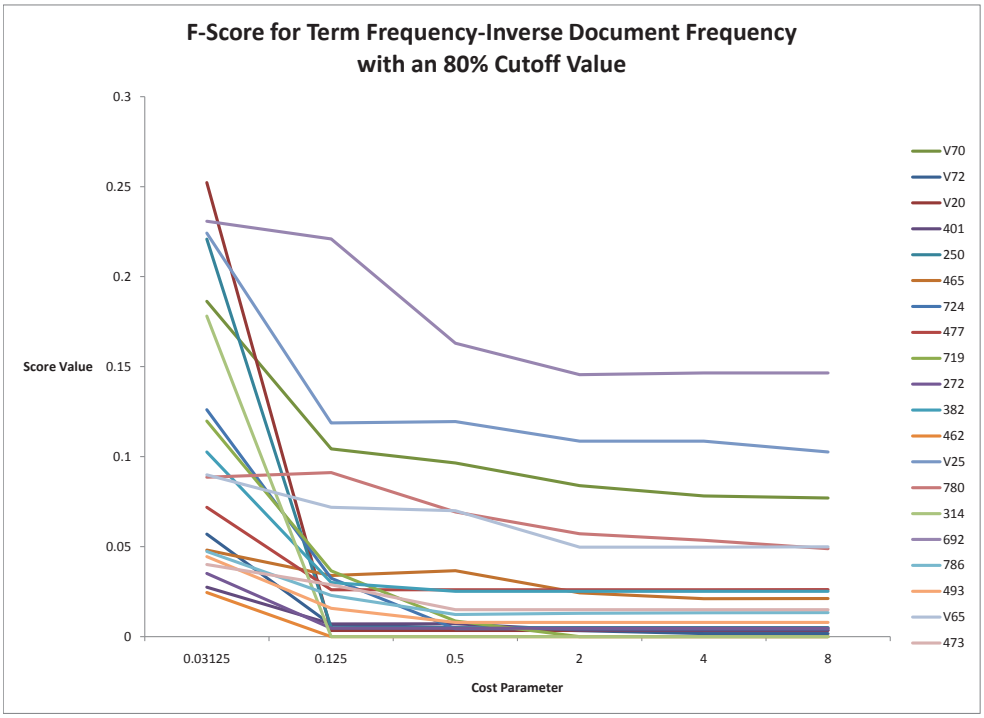
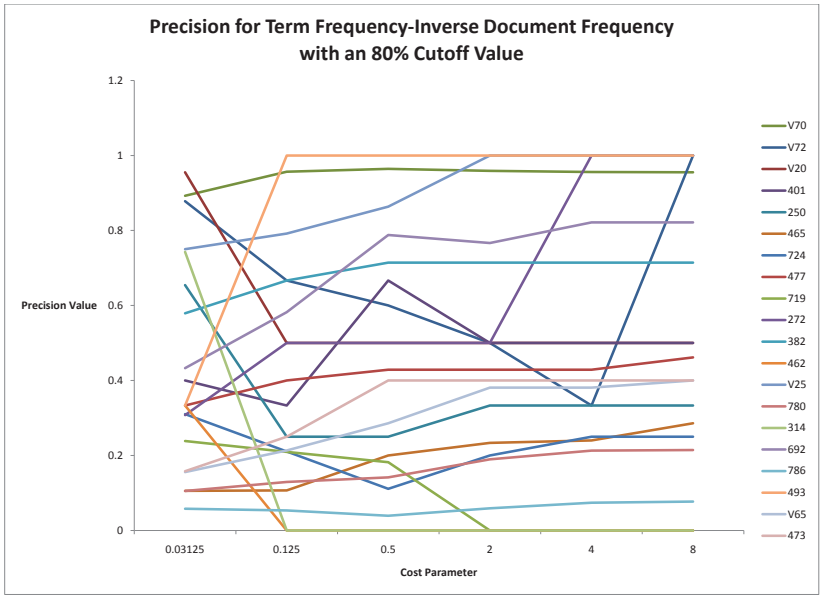


Figure 4.1: F-scores for the 80% cutoff threshold using the TfIdf weighting schema. We find that many of the codes have poor F-score values; however, a handful of scores may provide some promise. The F-score is the harmonic mean of the precision and the recall value. Therefore, if either the precision or the recall is lower than desired, we can tune the parameters to increase the F-score.

The initial results leave much to be desired; however, they are not completely unexpected and can be dealt with. The confidence cutoff was specifically used to improve the precision values with the foreknowledge that it could affect the recall values. As such, the cutoff parameter needs to be further tuned. We illustrate this tuning using a plot of the response surface for the F-score as we adjust the cost parameter and the confidence cutoff.

The response surface, presented in Figure 4.3, shows the effect of adjusting the cost parameter and the confidence cutoff. We find that reducing the cost parameter and the confidence cutoff provide the best results. V72, the code corresponding to Special Investigations, provides the best F-score amongst the 20 codes. Unfortunately, the F-score generated is 0.53, with a precision value of 0.61 and a recall value of 0.48.

The confusion matrix also provides insight into the applicability of this technique being imple-



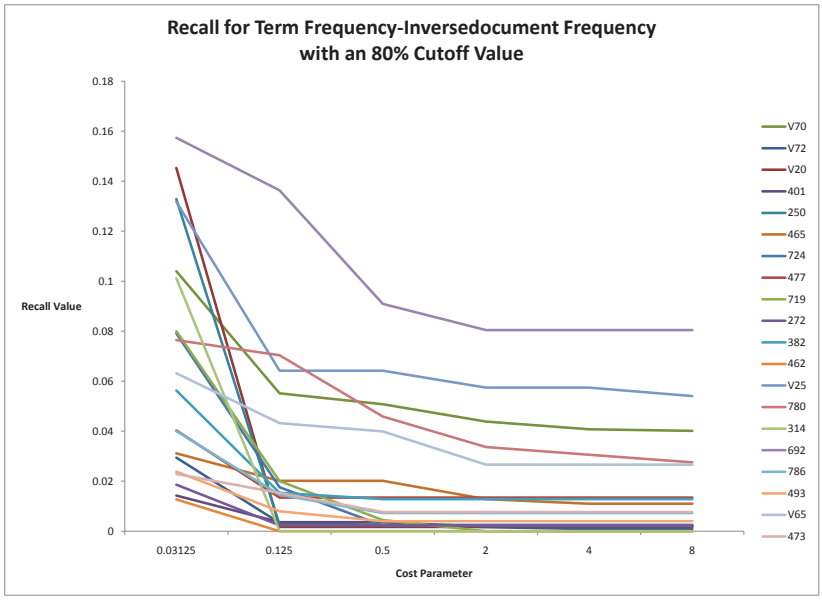


Figure 4.2: Precision and Recall for the 60% cutoff threshhold using the TfIdfN weighting schema. The precision value is the ratio of codes there were correctly predicted to the total number of codes predicted. While this chart shows a number of codes that oscillate at different values, we see that there are a handful of codes that have promise. We conclude that the recall values must be the issue with the current F-scores. The recall values leave much to be desired; however, it was not unexpected that the reacall values would be low. By implementing the confidence cutoff we enhanced the precision value and reduced the recall. By adjusting the confidence cutoff we should be able to ultimately improve the F-score.

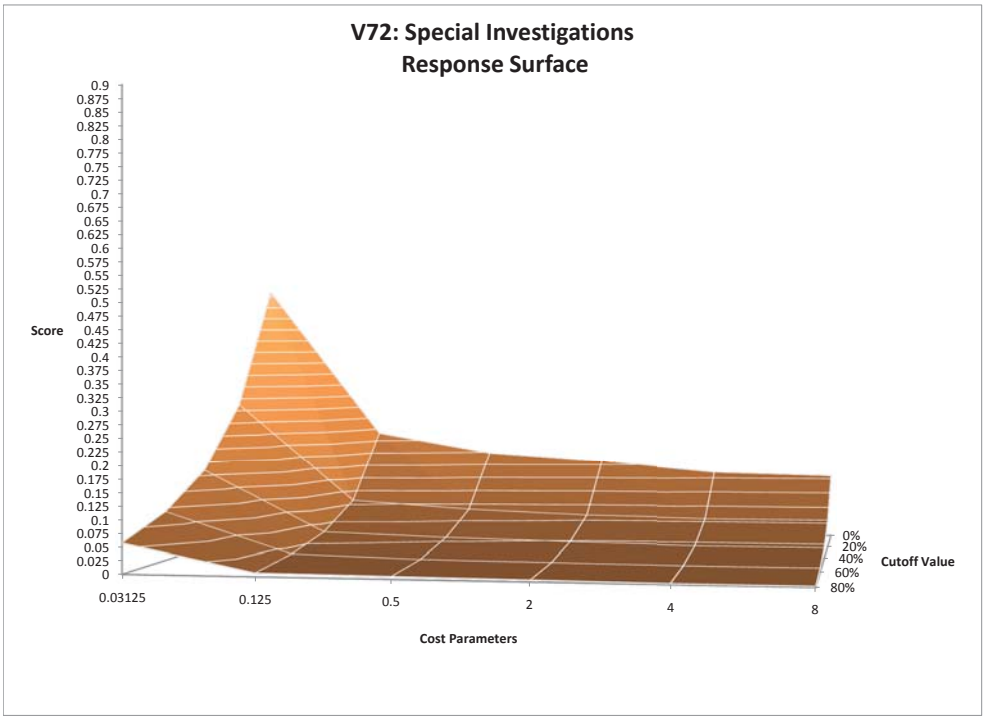


Figure 4.3: The Response Surface for code V72 using the TfIdf weighting schema. While the response surface indicates a significant increase in performance, we are still not much better off than we would have been flipping a coin. Perhaps further exploration into smaller cost values could improve the resulting F-score.

mented into industry. While we can be fairly confident of the accuracy of a prediction produced by the classifier, it only classifies a small portion of the total encounters correctly. For example, while the SVM makes 350 predictions, 279 of which were correct, we are still left with 725 that have not been classified. If the intention is to leverage statistical learning solely to enhance the accuracy of the codes we have, then the current SVM is a good potential candidate. If the intention is to reduce the human workload as well as to enhance the accuracy of the data, the SVM, in its current from, is not an appropriate tool.

From the response surface, we also learn that the term usage for this code is not necessarily linearly seperable. Evidence of this occurrence is presented by the F-score of 0 as the cost parameters trend to a hard margin classifier. In other words, the SVM provides better classification when terms are allowed to cross over the seperating hyperplane. Response surfaces for the remaining 19 codes can be found in Appendix B.

	Label y = V72	Label y = Unknown
prediction $\hat{y} = V72$	279	171
prediction $\hat{y}$ =Unknown	725	17,968

Table 4.1: Confusion Matrix for V72 using a cost value 2^{-5} and confidence cutoff 0. We find that the number of correct predictions is not satisfactory. While the SVM was fairly accurate when it made a prediction, it did not make enough predictions to actually be useful.

Term frequency-inverse document frequency weighting does not provide the type of results we expected from our initial research. In fact, the macro-average score for this weighting schema was a dismal 0.26 with a range of 0.53 to 0.05. It is not readily apparent why this schema did not classify more appropriately. The second methodology, normalized term frequency-inverse document frequency, should provide a better classifier given the normalization component of the weighting.

4.1.2 Normalized Frequency-Inverse Document Frequency

Normalized frequency-inverse document frequency provides a slightly more sophisticated weighting schema than term frequency-inverse document frequency. While the normalized version provides the ability to normalize documents that are written by authors who utilize a more verbose dictionary. In other words, it puts documents that contain many terms on the same playing field as documents that contain very few terms.

As with term frequency-inverse document frequency schema, we begin our analysis by reviewing the performance of each model using six different cost parameters with a confidence cutoff of 60%. However, Figure 4.4 shows the same trends in F-score values as seen in Figure 4.1. As before, precision and recall values are broken out to determine if the recall is the greatest hindrance to the success of this classifier.

Figure 4.5 shows that the term frequency-inverse document frequency plots resemble those of Figure 4.1; however, the recall has improved slightly. Again a handful of codes have promising precision values that appear to be degraded by the poor recall values. Altering the tuning parameters shows promise with the previous data set and are attempted as well. We employ response surfaces to visualize the effect of changing the tuning parameters has on the F-score.

Altering the tuning parameters, does provide some benefit; however, it still does not strengthen the classifier enough to be a viable option. The tuning parameter confidence cutoff tested the values 0.8, 0.6, 0.4, 0.2 and 0.0 as is represented in Figure 4.6. The cost parameter were also

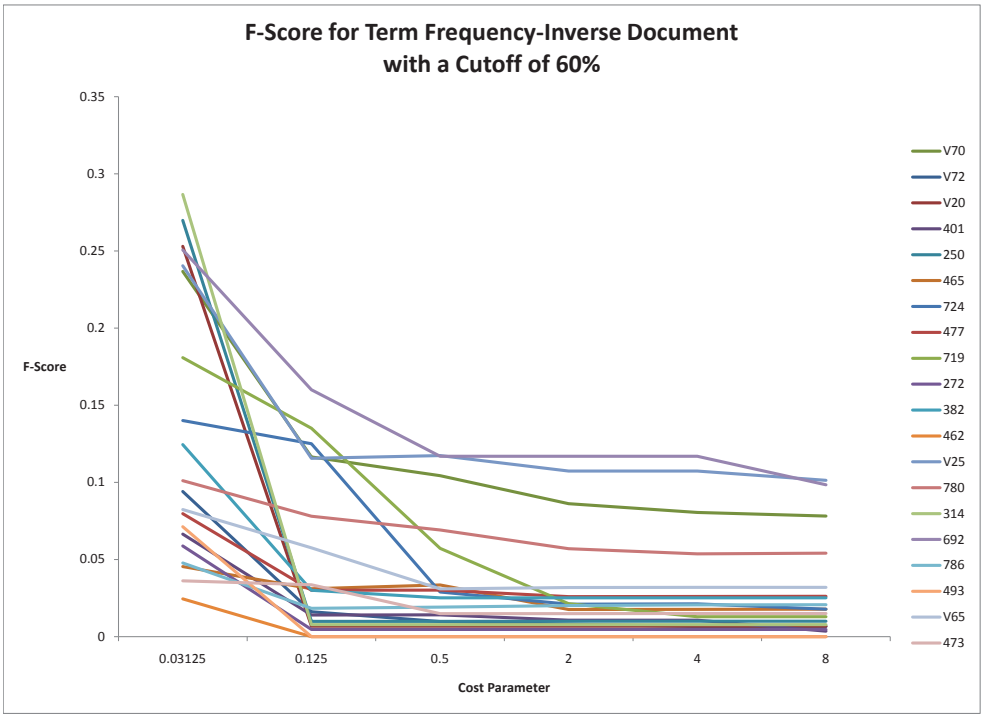
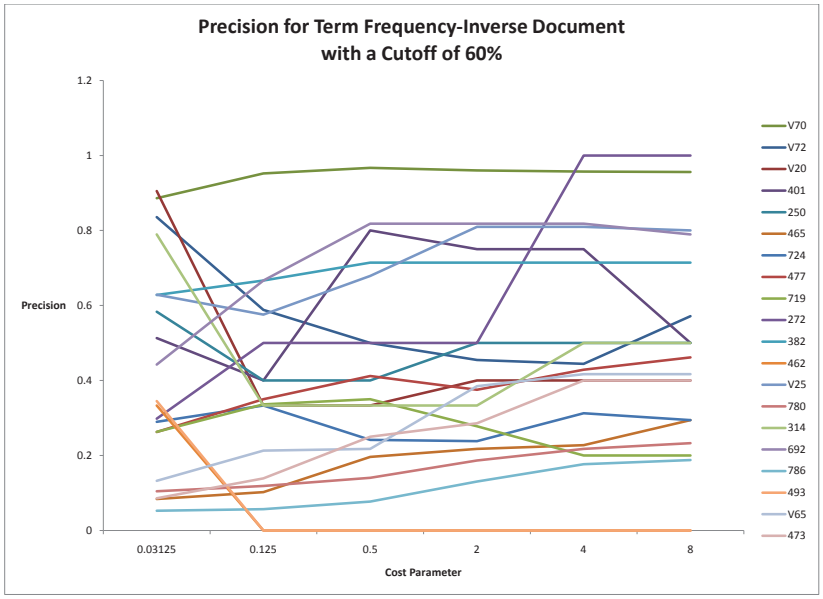


Figure 4.4: F-scores for the 60% cutoff threshold using the TfIdfN weighting schema. The F-scores are not satisfactory and require further investigation. Again, breaking the F-score into its parts, precision and recall, should provide the direction required to improve the overall performance of the classifier.

tested in the set C described earlier. We see the affect of the changes in the paramters in the response surfaces. Again code V72 presents the best classified code and will be used to illustrate the performance. The full list of response surfaces can reviewed in Appendix C.

The response surface corresponding to the normalized term frequency-inverse document frequency for code V72, presented in Figure 4.6, is similar to the non-normalized version. The precision value is 0.62, a 0.01 increase from the non-normalized dataset, and the recall is 0.47, a 0.01 decrease from the non-normalized dataset, resulting in the F-score remaining constant at 0.53.

Code V72 only represents one code across the classification spectrum and so we must calculate the macro-average F-score in order to properly compare the two weighting techniques employed thus far across all twenty diagnosis codes. We find that the macro average F-score for



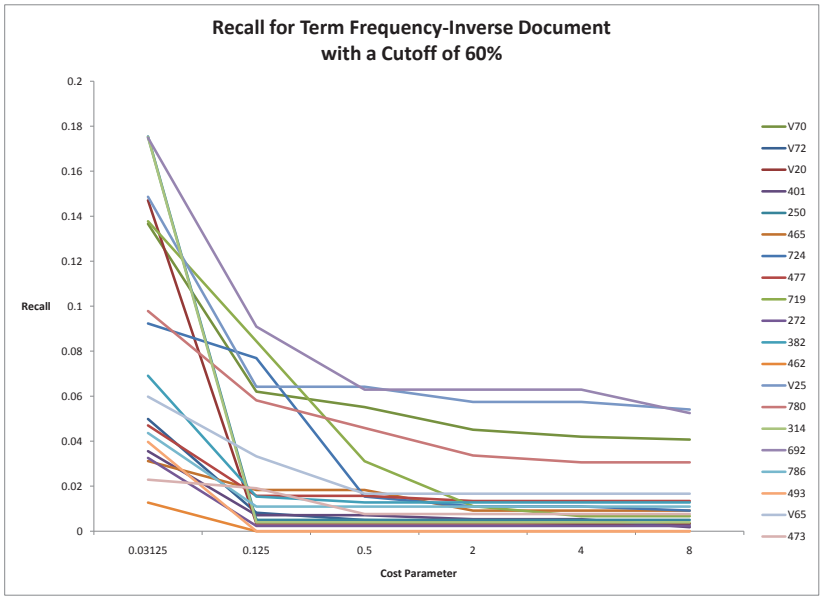


Figure 4.5: Precision and Recall for the 60% cutoff threshhold using the TfIdfN weighting schema. The precision value is the ratio of codes there were correctly predicted to the total number of codes predicted. While this chart shows a number of codes that oscillate at different values, we see that there are a handful of codes that have promise. We conclude that the recall values must be the issue with the current F-scores. The recall values leave much to be desired; however, it was not unexpected that the reacall values would be low. By implementing the confidence cutoof we enhanced the precision value and reduced the recall. By adjusting the confidence cutoff we should be able to ultimately improve the F-score.

the normalized data set is 0.25 with a range of 0.53 to 0.06. Therefore, we can conlclude that for this particular data set, normalizing is a hindrance to the classification process.

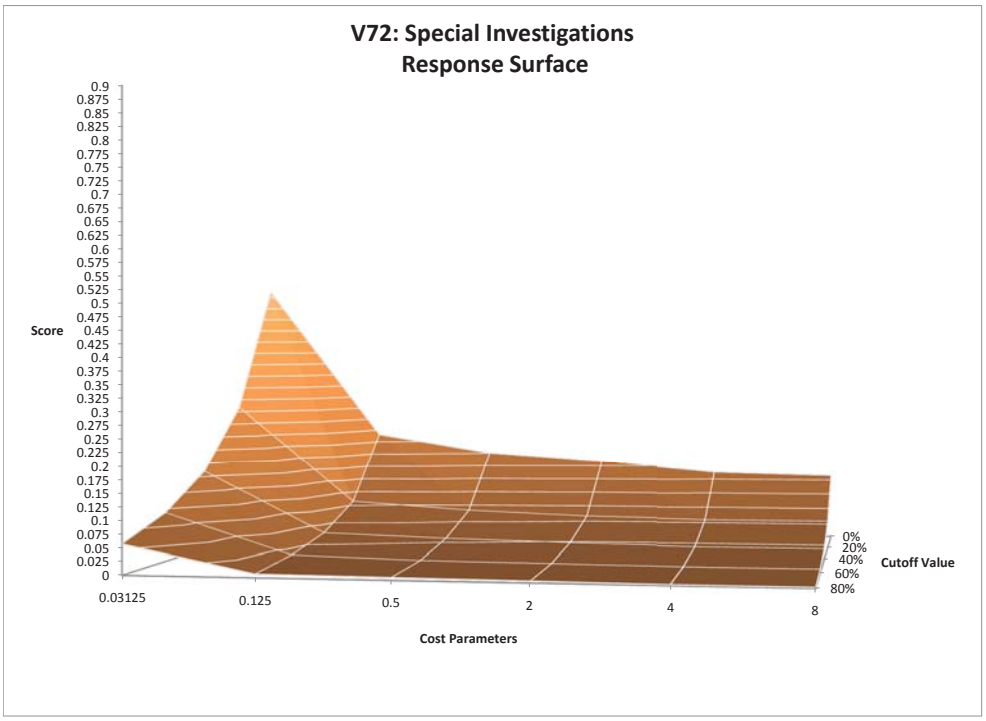


Figure 4.6: The Response Surface for code V72 using the TfIdfN weighting schema. While the response surface indicates a significant increase in performance, we are still not much better off than we would have been flipping a coin. Perhaps further exploration into smaller cost values could improve the resulting F-score.

4.1.3 Term Frequency

Term Frequency is the base weighting schema used in this thesis and will serve as the final weighting schema to be analyzed. As a reminder, the term frequency matrix is composed of the counts of each term for each document in the corpus. While this is the least sophisticated weighting scheme it has been shown to provide sucessful results in other contexts. [3]

We begin our analysis in the same manner as before, an inspection of the F-score with a rigorous confidence cutoff of 0.8. Figure 4.7 shows the twenty codes selected for this analysis and their corresponding F-score at the 80% cutoff. These scores are significantly better than the two previous methodologies. Unfortunately, no code's F-score meets the goal F-score of 0.8 at the 80% confidence cutoff requiring further tuning of the paramters.

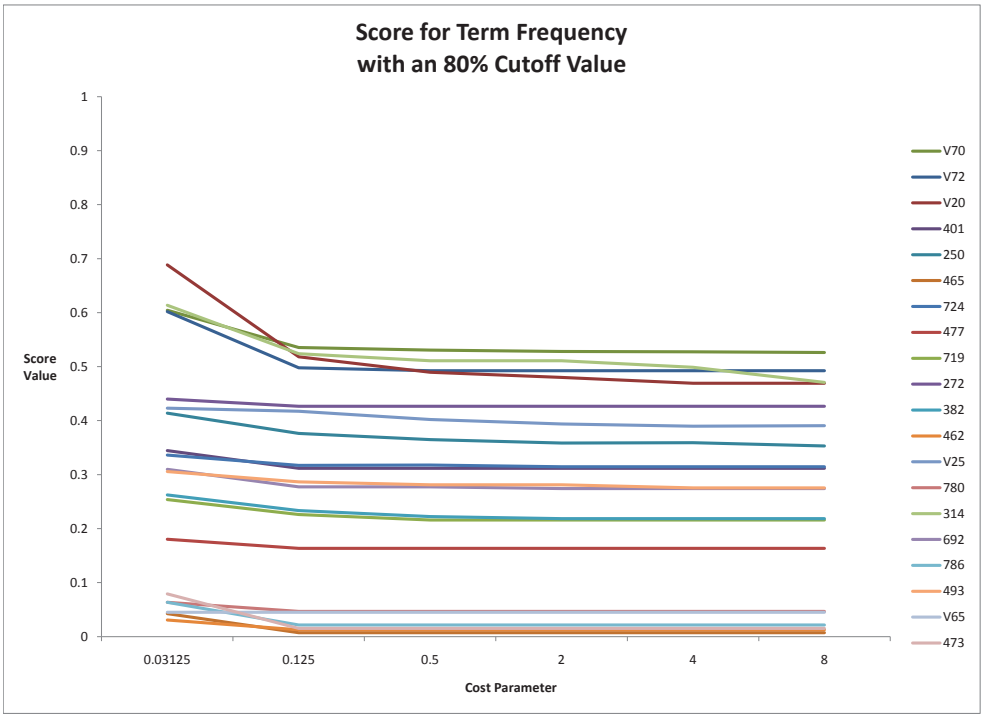
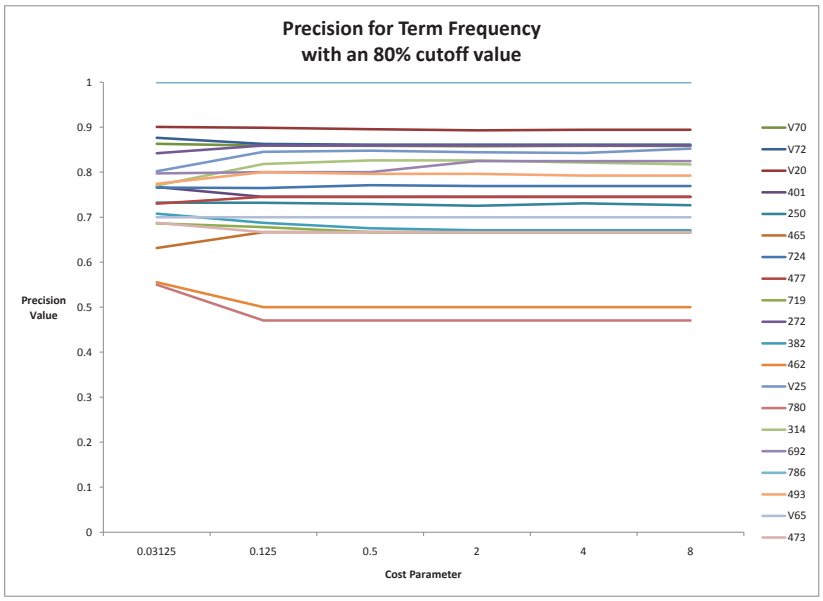


Figure 4.7: F-scores for the 80% cutoff threshold

With only a handful of codes currently near the goal of 0.8, we must refine our classification methodology. Figure 4.8 shows that the precision of the model is quite good for most of the records. This is not surprising as we have artificially increased this metric by requiring the confidence of the SVM to be above a very high threshold. Again, we conclude that the recall for these models must be at fault.

Figure 4.8 presents recall values across the six cost parameters. The best recall metrics are below 0.6 and deteriorate quickly. It is readily apparent that the cost variable directly corresponds to the recall and precision metric and hopefully can be enhanced by continuing to tune the confidence cutoff value. Visualizing the response surface created by the cost variable and the confidence cutoff value should aid in selecting the optimal combination. With such positive performance from this classifier thus far a series of response surfaces will be created for the four codes that have the most potential of becoming successfully classified: V70,V72, V20, and 314. We delve deeper into the response surfaces of each of these codes as we vary the tuning



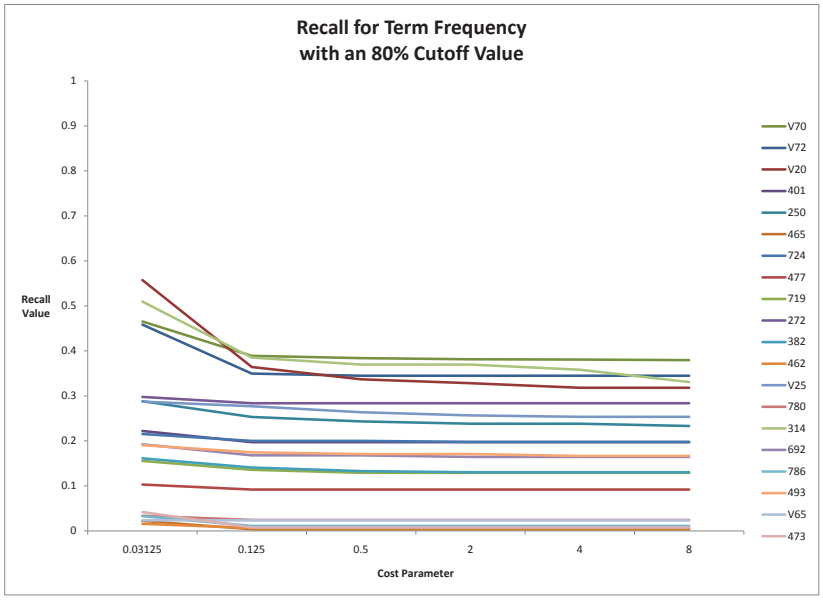


Figure 4.8: Precision and Recall for the 80% cutoff threshhold using the Term Frequency weighting schema. The precision value is the ratio of codes that were correctly predicted to the total number of codes predicted. The worst precision value for the Term Frequency schema is significantly better than the best values of the two more complicated weighting schemas. Furthermore, we find that the recall values are still the limiting factor, but, as we have shown before, we can adjust the confidence cutoff in order to enhance the F-score.

parameters. The initial response surface for V70 can be seen in Figure 4.9, which presents the F-scores at each of the parameter values.

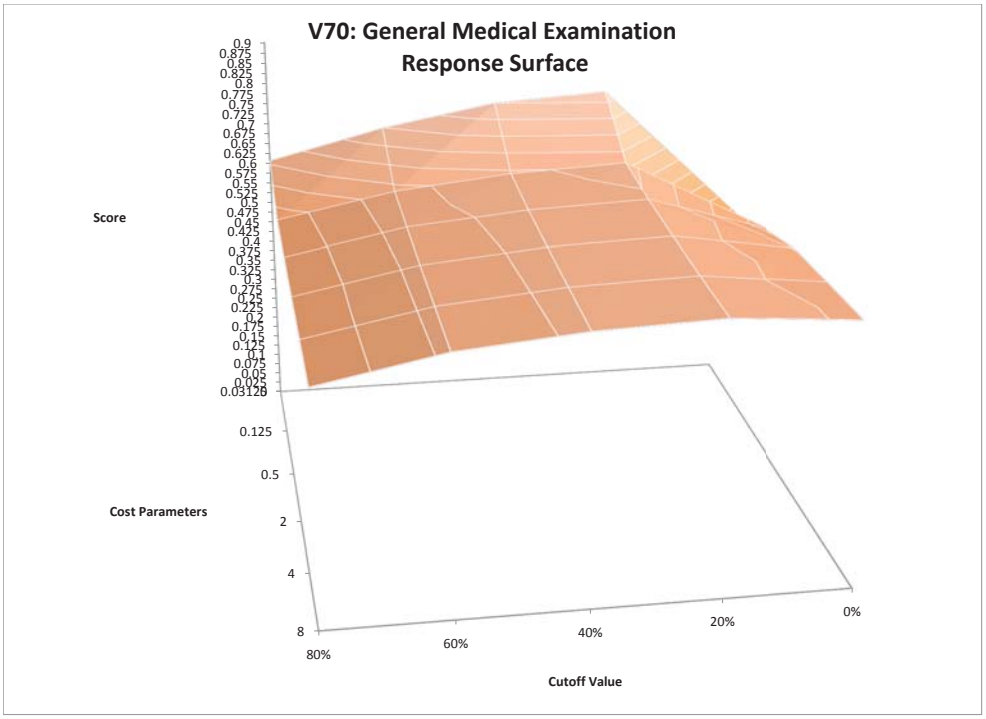


Figure 4.9: Response Surface for Code V70: General Medical Examination. The response surface indicates that the F-score greatly increases as the cost variable value is decreased. Furthermore, we find a peak at the confidence cutoff level of 20%. If this trend continues, further examination into smaller cost variables will be conducted to determine if the F-score goal can be exceeded.

Figure 4.9 illustrates the relationship between the two tuning parameters, cost and confidence cutoff, while plotting the corresponding F-score value. The resulting F-score increases as the confidence cutoff and the cost variables are decreased. We find that the smallest cost parameter, 2^{-5}, and a confidence cutoff value of 20% provide the best results. An F-score of 0.74, comprised of a precision value of 0.76 and a recall value of 0.73, is the best classification mix that can be achieved for code V70 under the current spread of parameters. Unfortunately, in the case of V70 we never exceed the 0.8 threshold deemed a successful autoclassification. The trend does indicate that building models with smaller values of the cost parameter may produce satisfactory results. We investigate if this trend exists throughout the remaining codes.

Figure 4.10 demonstrates the F-score values for the code V72: Special Investigations. Again, a sharp increase in F-score by the response surface indicates that smaller cost variable values and a confidence cutoff value of 20% provide the best combination of tuning parameters. V72's ultimate F-score is 0.75 with a precision value of 0.8 and a recall value of 0.72. Both of these codes are very close to being successfully classified under our initial definition of success. The same behavior is manifested throughout the remaining codes (see Appendix A) resulting in a macro-averaged score is 0.51 with a range of 0.76 to 0.11. Not surprisingly, the term frequency weighting schema produces the best macro-averaged F-score among the three weighting schemas tested.

Figure 4.10: V72 Response Surface

We have shown that the term frequency weighting schema is the most promising in our initial pass through the data and we must now determine where to focus our research. While 80% of the codes, 16 out of 20, report the highest F-score values with a cost parameter of 2^{-5} and a cutoff value of 20% we are not ready to ultimately declare that tuning parameter combination our best model. Figure 4.11 shows each code and its respective F-scores by the two smallest

cost values used, the values on the x axis, as well as cutoff values, the colored bars. As you can see, the 20% cutoff bar, in orange, eminating from the 2^{-5} cost value, is most often the largest F-score value in the group. In rare cases we find that the cutoff value of 0, in blue, has the best F-score. The cases where the cutoff value of 0 is chosen tend not to categorize well, evidenced by the best F-score remaining below 0.4, and provides more proof that the best combination of tuning parameters is a cost value of 2^{-5} and a confidence cutoff of 20%.

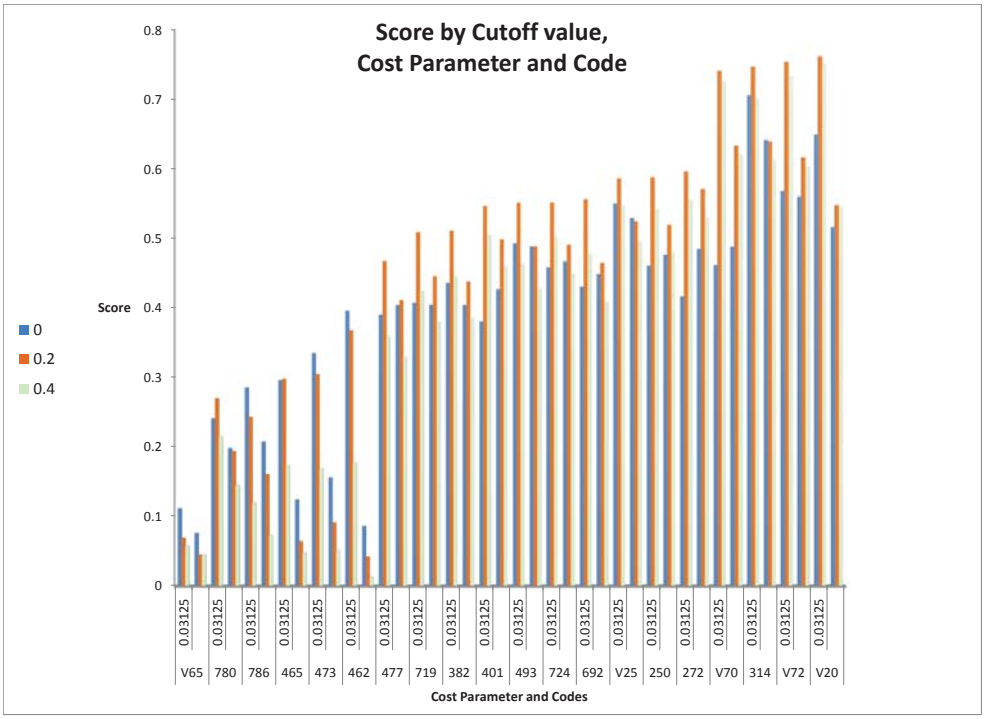


Figure 4.11: The F-scores of all 20 codes by Cutoff Value and cost Value

In regards to the term frequency weighting, under the current parameter values we conclude that the best combination of parameters, to be referred to as the candidate model, is a confidence cutoff value of 20% and a cost variable value of 2^{-5}; however, searching the response surface beyond cost values of 2^{-5} may yield more promising results.

4.2 Final Model

We conclude our discussion of the model formulation with the presentation of the results from applying the candidate model to the test data set. The test set was prepared in the same fashion as the training set. We find that the test set does not classify as well as the validation set. In fact, the classification is quite poor.

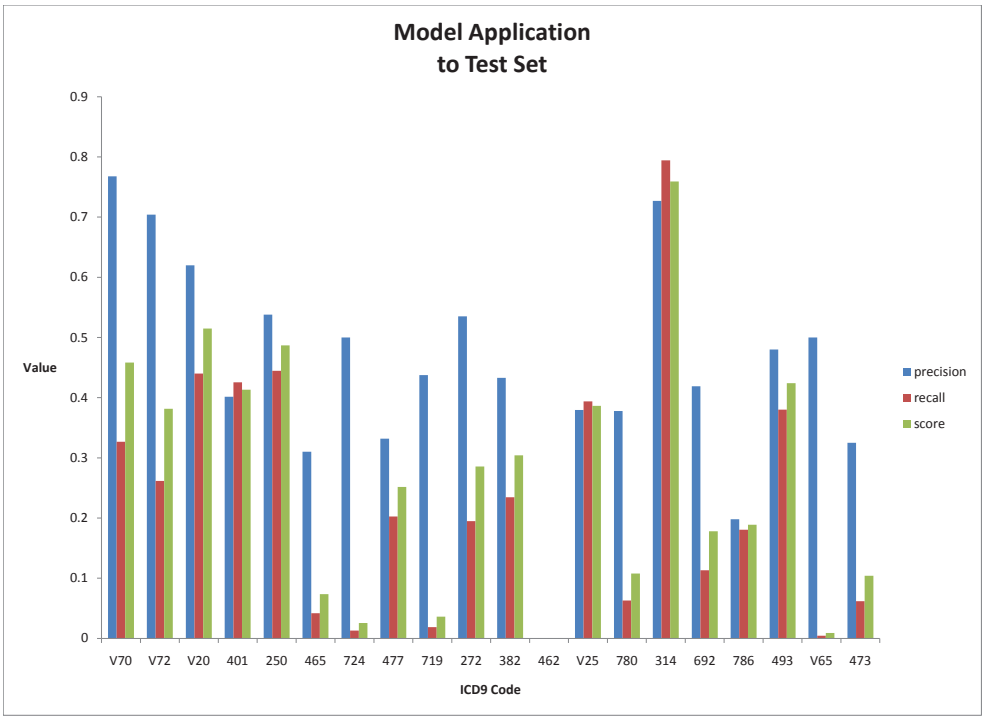


Figure 4.12: Results from the application of the candidate model to the test set of data. We see that the results are not as positive as the results discovered using the validation set.

We find that the precision value remains similar to the values derived in the validation stage of research. The recall values still remain poor which affects the overall score. While this was mitigated in the validation stage of the experiment, further research needs to be conducted in order to find methods that are more robust in regards to the recall value. Figure 4.12 shows the precision, in blue, recall, in red, and score, in green, for the final model. Only one code, 314 Hyperkinetic Disorders Childhood, classifies above the successful threshold using the candidate model. The model, in its current form, does not provide enough discrimination between codes to function as a stand-alone coding classifier. Further research must be conducted in order to

provide an autoclassification technique that can be utilized by the MHS.

CHAPTER 5:
Conclusion

Healthcare delivery is in need of technological injections that provide value to clinicians while enhancing the efficacy of the documentation system. This thesis lays the foundation for future work in classification of ICD9 coding for MHS. The greatest and first hurdle to such work is the arduous and time-consuming preperation of the physician's notes into a corpus of useable documents and terms and their subsequent translation into TDMs. In this thesis, we provide a process consisting of a series of programs which convert TDMs to more memory and analytically efficient sparse matrix formats which can be used for classification.

In this work we used support vector machine classifiers which have proven successful in similiar classification of text in medical contexts. [3, 5, 7] Interestingly, of the three weighting schema investigated, the simplest technique, term frequency, perfomed the best. This result contrasts with the results of [3] which finds that the more complicated schema; term frequency-inverse document frequency and normalized inverse document frequency, tend to perform better. Further work will need to be done to determine the reasons for methods, that have proven successful previously, to perform so poorly on MHS data. We suspect that the most common terms may have been due to AHLTA's auto-population resulting in a bulk of common words throughout all of the records. Furthermore, the culture surrounding AHLTA's utilization may negate some of the traditional documentation practices that are common in civilian practice. A final area of weakness in the data could be from the coding of the records used in the training set. Data, in which physicians assign the ICD9 code, was used in order to prove the concept without large financial expenditure. It may be that a clean set of professionally coded data is required to provide the discriminatory power required. In addition to these data issues, we suggest changes to the model that include exploring cost values less than 2^{-5} as well as more refined increments of cutoff values. A full list of potential areas of further analysis include

- exploring cost values less than 2^{-5} (previously mentioned)
- Smaller increments of confidence cutoff values (previously mentioned)
- Adding extenders to the codes

- Analysis of the term frequency-inverse document frequency and normalized term frequency-inverse document frequency matrices to determine why the classification of these methods was so poor. (previously metioned)
- Using the probabilitiy outputs of the SVM as a meta-model to be fed into a classification tree that also utilizes the demographic information
- Additional classification engines such as naïve bayes, and k-nearest neighbors

We focused solely on the physician text and how a SVM could descriminate between codes. Other classifiers, which have performed well in classification of text in a medical context, are available for application. These techniques include the naïve bayes classifier and the k-nearest neighbor classifier. Even more important, in this thesis we focus on the physician notes to classify codes. While the notes provide the most pertinent and detailed information for diagnosis, other data, availabe from the CDM, can also be used for good purpose. For example, we might partition the data by age and gender prior to developing classification schemes. Furthermore, fields such as appointment type, Family Member Prefix, and MEPRS code can be incorporated directly into the physician note SVM classifiers. With additional data processing, patient histories can be constructed from CDM data. These might also prove valuable for classification. Alternatively, because constructing SVM classifiers based on physician notes is so time-consuming and labor-intensive, the SVM classifiers might be used as a first stage in classification. The output probabilities of the candidate diagnosis code from the SVM text classifier, along with demographic information available from CDM records, could be used as more manageable input to a second stage classifier. This approach to statisical learning is call "chaining" and is described in [2].

Finally, to be useable, this work must be extended to more diagnosis codes and to include the code extenders. Adding extenders may uncover the unique vector space that comprises each of these codes. Finding this smaller space will require enhanced computational assets, but may provide a better vector space in which to classify. More importantly, adding the extenders will provide an output that is practical and necessary for the leadership of the MHS. Continued research in these areas may provide the necessary boost to SVM model to provide an ICD9 autoclassification engine. ICD9 autoclassification is an important first step in providing objective data for the MHS leadership and deserves further research.

LIST OF REFERENCES

[1] Leiyu Shi and Douglas A. Singh. *Delivering Health Care in America, A Systems Approach*. Prentice Hall International, 1997. ISBN 0-7637-3199-4.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009. ISBN 978-0-387-84857-0.

[3] T. Joachims. A statistical learning model of text classification with support vector machines. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 128–136, 2001.

[4] George Forman. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3:1289–1305, March 2003. ISSN 1532-4435. http://portal.acm.org/citation.cfm?id=944919.944974.

[5] Leah S. Larkey and W. Bruce Croft. Combining classifiers in text categorization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pp. 289–297. ACM, New York, NY, USA, 1996. ISBN 0-89791-792-8. http://doi.acm.org/10.1145/243199.243276.

[6] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pp. 412–420. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997. ISBN 1-55860-486-3. http://portal.acm.org/citation.cfm?id=645526.657137.

[7] John P. Pestian, Christopher Brew, PawełMatykiewicz, D. J. Hovermale, Neil Johnson, K. Bretonnel Cohen, and Duch. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, BioNLP '07, pp. 97–104. Association for Computational Linguistics, Stroudsburg, PA, USA, 2007. http://portal.acm.org/citation.cfm?id=1572392.1572411.

[8] Yi Zhang and Bing Liu. Semantic text classification of disease reporting. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pp. 747–748. ACM, New York, NY, USA, 2007. ISBN 978-1-59593-597-7. http://doi.acm.org/10.1145/1277741.1277889.

[9] Ingo Feinerer. *tm: Text Mining Package*, 2011. http://tm.r-forge.r-project.org/. R package version 0.5-6.

[10] Duncan Temple Lang. r-cran-xml, 2011.

[11] Ingo Feinerer, Kurt Hornik, and David Meyer. Text mining infrastructure in r. *Journal of Statistical Software*, 25(5):1–54, 3 31, 2008. CODEN JSSOBK. ISSN 1548-7660. http://www.jstatsoft.org/v25/i05.

[12] Evgenia Dimitriadou, Kurt Hornik, Friedrich Leisch, David Meyer, and Andreas Weingessel. r-cran-e1071, 2011.

[13] George K. Zipf. *Human Behavior and the Principle of Least Effort*. 1949.

[14] Gerard Salton and Chistopher Buckley. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing Management*, 24(5):513–523, November 1988.

[15] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In *Machine Learning*, pp. 273–297, 1995.

APPENDIX A:
Additional Term Frequency Response Surfaces

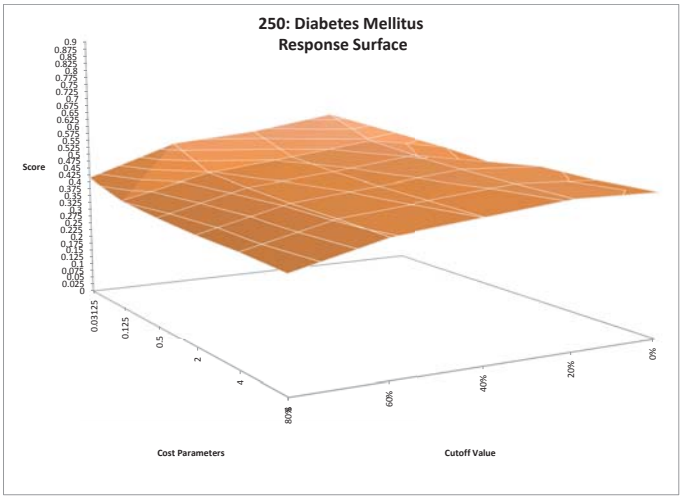


Figure A.1: The F-score response surface of 250 created by the tuning of the cost parameter and the confidence cutoff for term frequency weighting

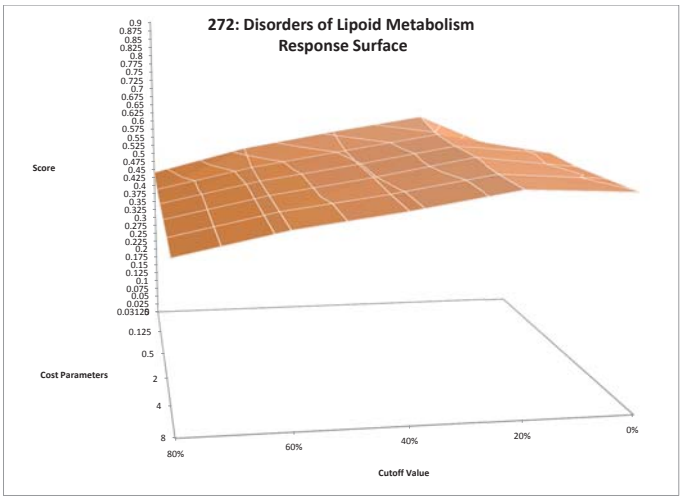


Figure A.2: The F-score response surface of 272 created by the tuning of the cost parameter and the confidence cutoff for term frequency weighting

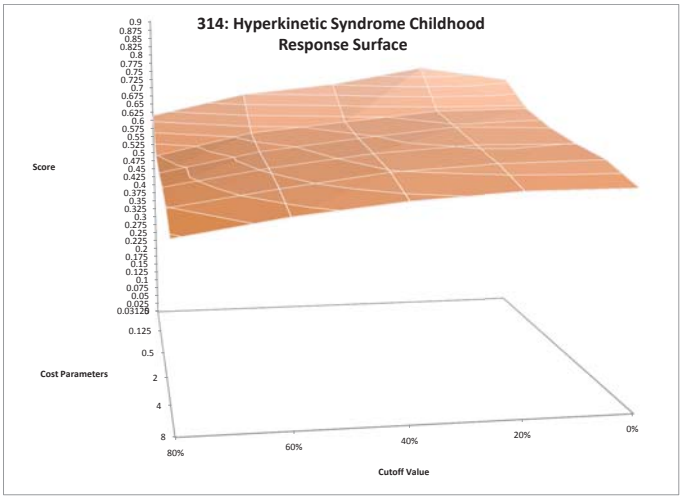


Figure A.3: The F-score response surface of 314 created by the tuning of the cost parameter and the confidence cutoff for term frequency weighting

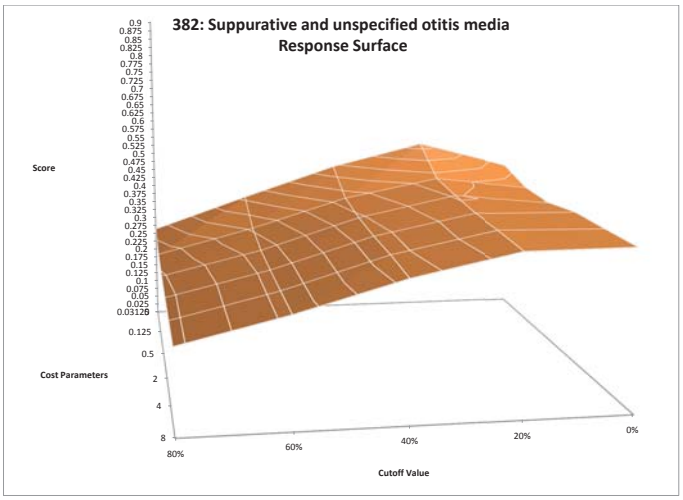


Figure A.4: The F-score response surface of 382 created by the tuning of the cost parameter and the confidence cutoff for term frequency weighting

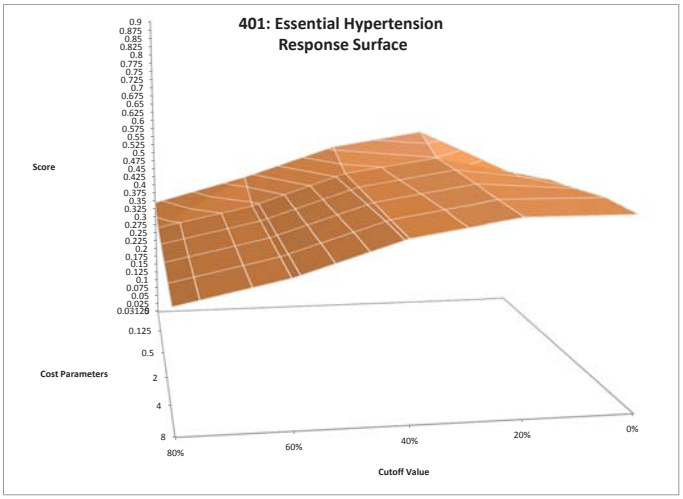


Figure A.5: The F-score response surface of 401 created by the tuning of the cost parameter and the confidence cutoff for term frequency weighting

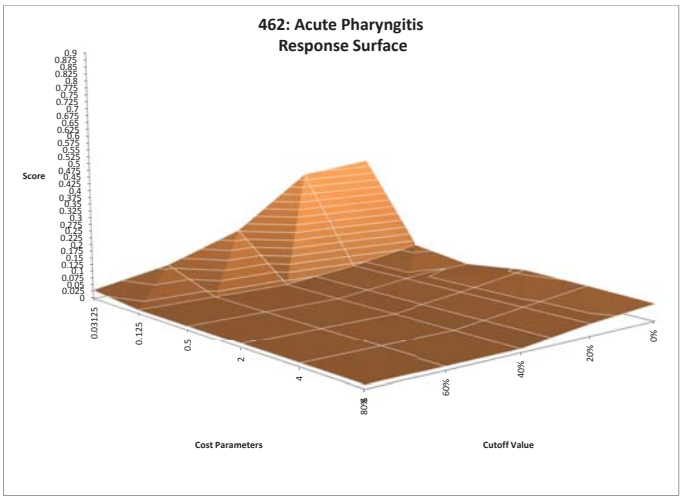


Figure A.6: The F-score response surface of 462 created by the tuning of the cost parameter and the confidence cutoff for term frequency weighting

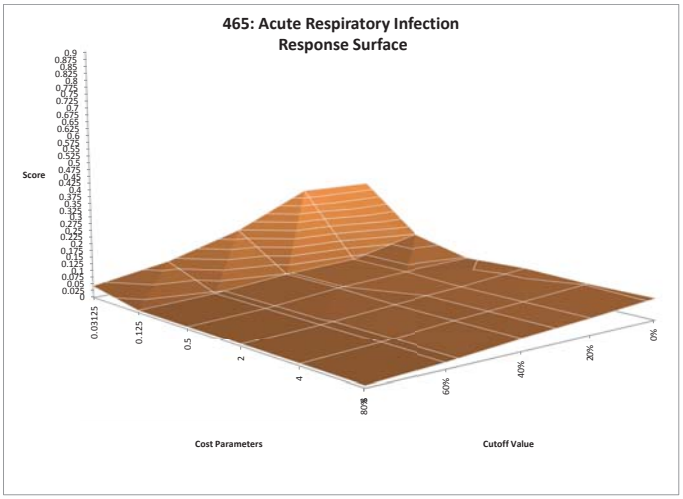


Figure A.7: The F-score response surface of 465 created by the tuning of the cost parameter and the confidence cutoff for term frequency weighting

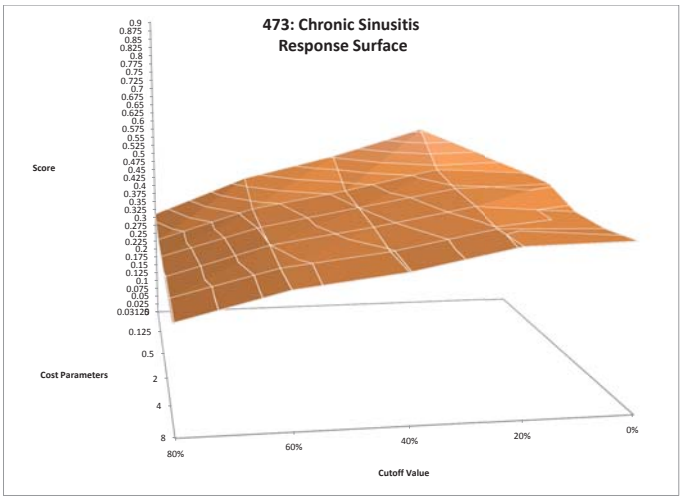


Figure A.8: The F-score response surface of 473 created by the tuning of the cost parameter and the confidence cutoff for term frequency weighting

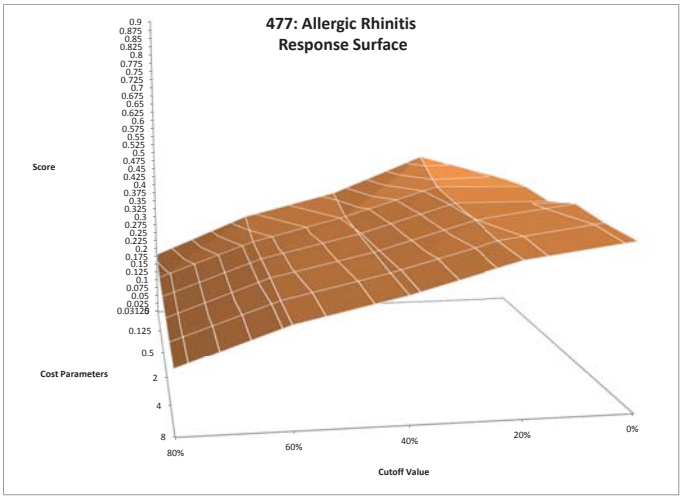


Figure A.9: The F-score response surface of 477 created by the tuning of the cost parameter and the confidence cutoff for term frequency weighting

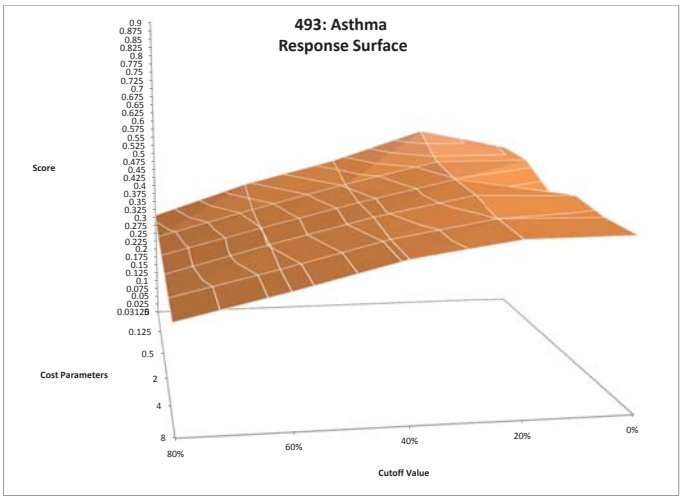


Figure A.10: The F-score response surface of 493 created by the tuning of the cost parameter and the confidence cutoff for term frequency weighting

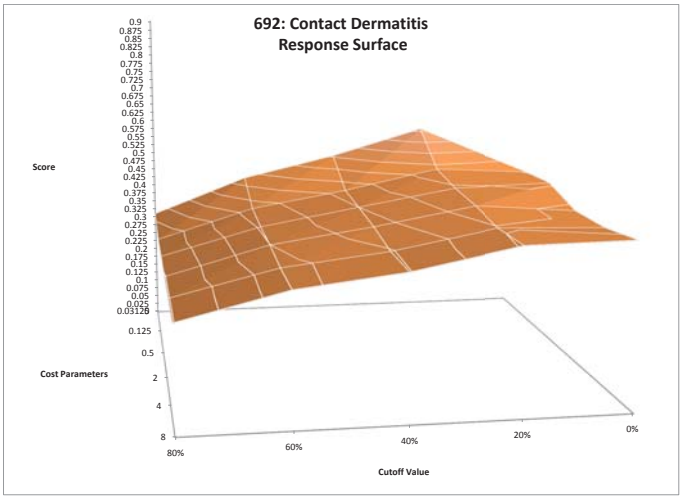


Figure A.11: The F-score response surface of 692 created by the tuning of the cost parameter and the confidence cutoff for term frequency weighting

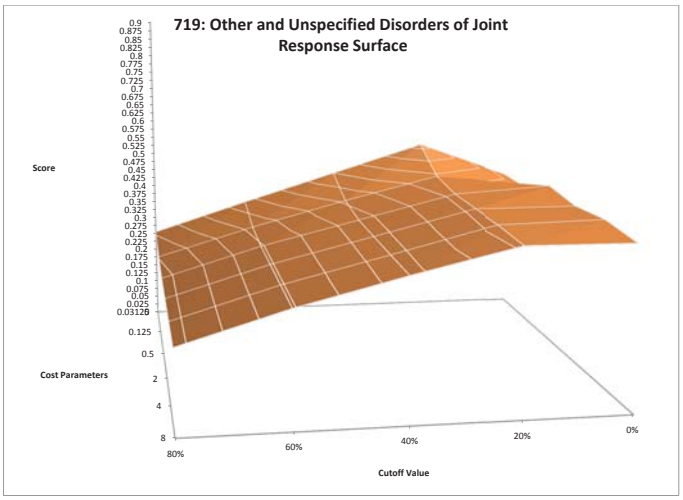


Figure A.12: The F-score response surface of 719 created by the tuning of the cost parameter and the confidence cutoff for term frequency weighting

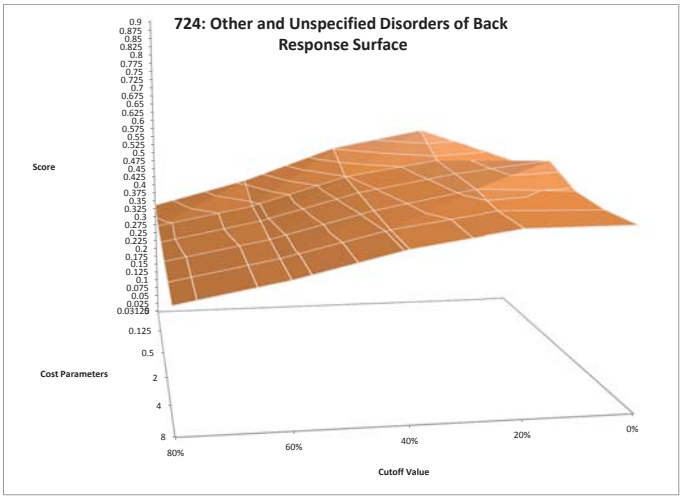


Figure A.13: The F-score response surface of 724 created by the tuning of the cost parameter and the confidence cutoff for term frequency weighting

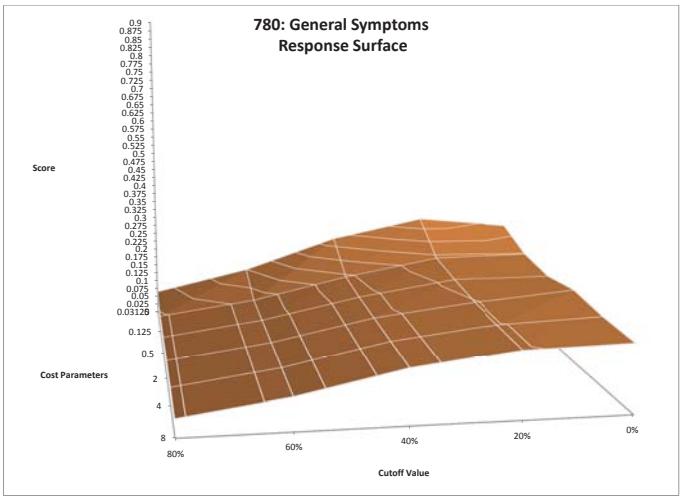


Figure A.14: The F-score response surface of 780 created by the tuning of the cost parameter and the confidence cutoff for term frequency weighting

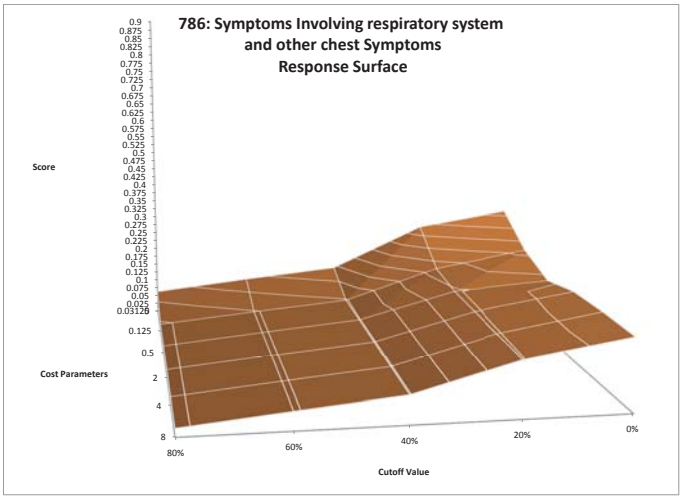


Figure A.15: The F-score response surface of 786 created by the tuning of the slack variable and the confidence cutoff for term frequency weighting

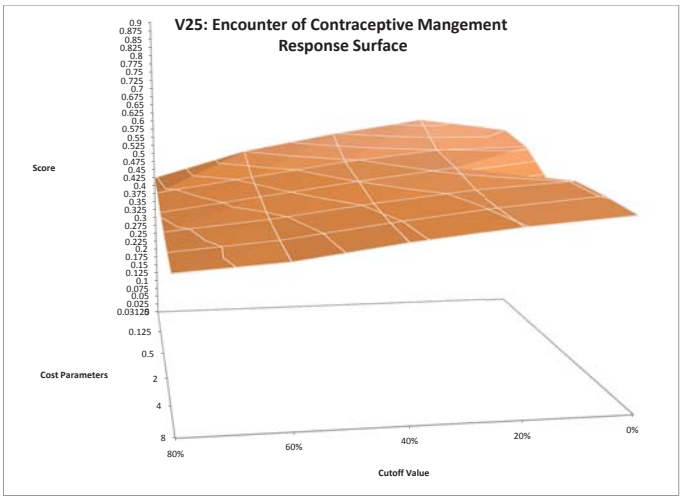


Figure A.16: The F-score response surface of V25 created by the tuning of the cost parameter and the confidence cutoff for term frequency weighting

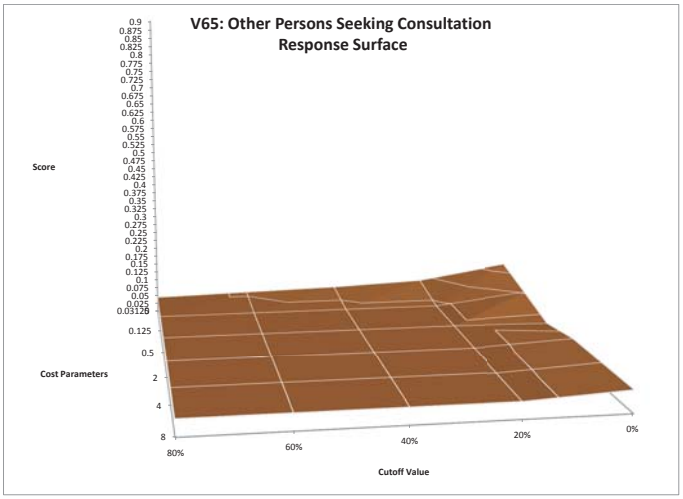


Figure A.17: The F-score response surface of V65 created by the tuning of the cost parameter and the confidence cutoff for term frequency weighting

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B:
Additional Term Frequency-Inverse Document Frequency ResponseSurfaces

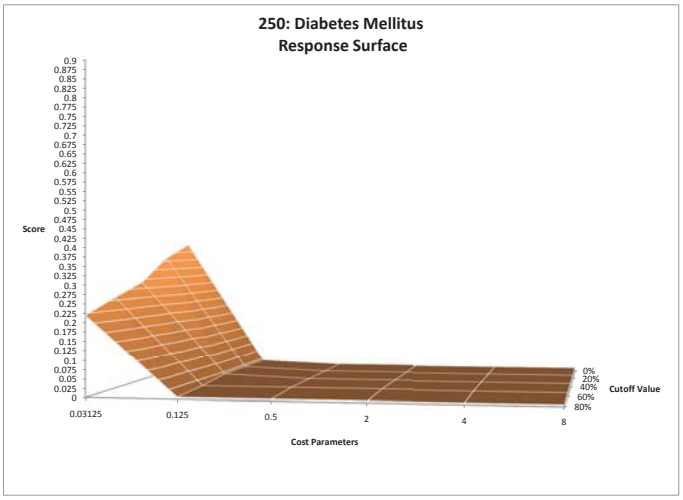


Figure B.1: The F-score response surface of 250 created by the tuning of the cost parameter and the confidence cutoff for term frequency-inverse document frequency weighting

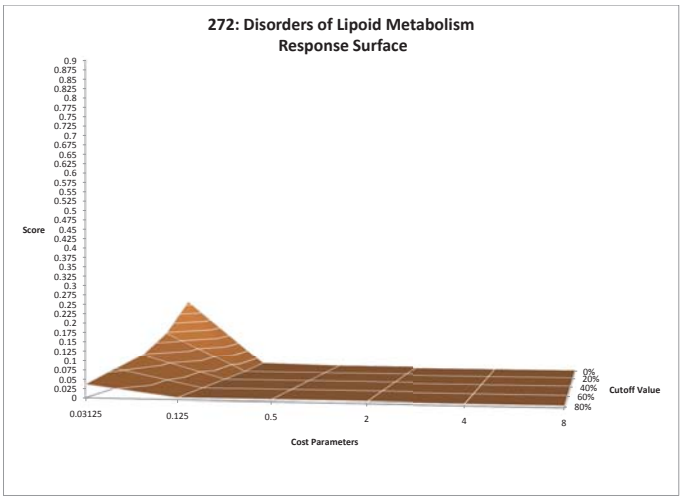


Figure B.2: The F-score response surface of 272 created by the tuning of the cost parameter and the confidence cutoff for term frequency-inverse document frequency weighting

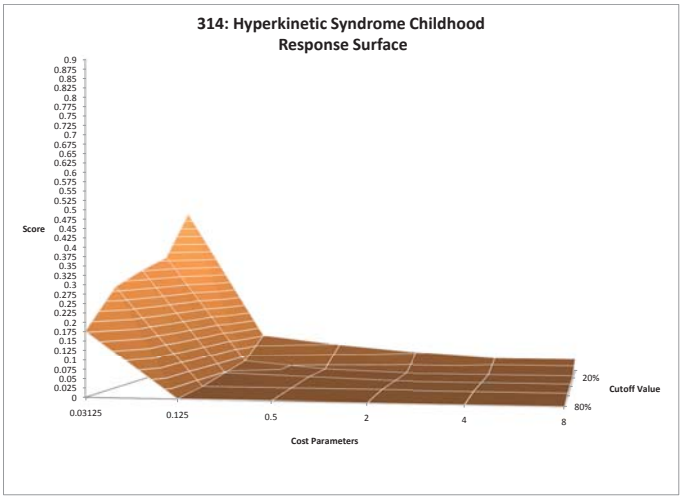


Figure B.3: The F-score response surface of 314 created by the tuning of the cost parameter and the confidence cutoff for term frequency-inverse document frequency weighting

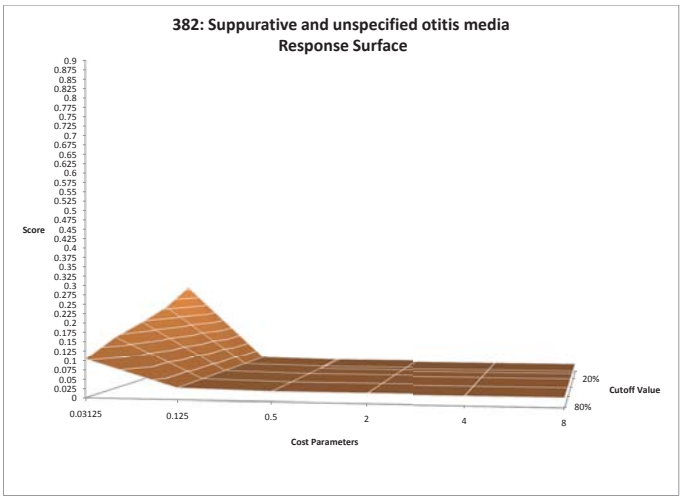


Figure B.4: The F-score response surface of 382 created by the tuning of the cost parameter and the confidence cutoff for term frequency-inverse document frequency weighting

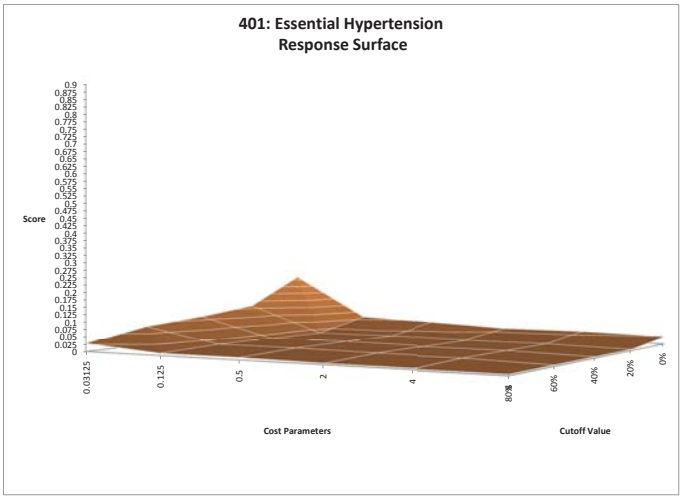


Figure B.5: The F-score response surface of 401 created by the tuning of the cost parameter and the confidence cutoff for term frequency-inverse document frequency weighting

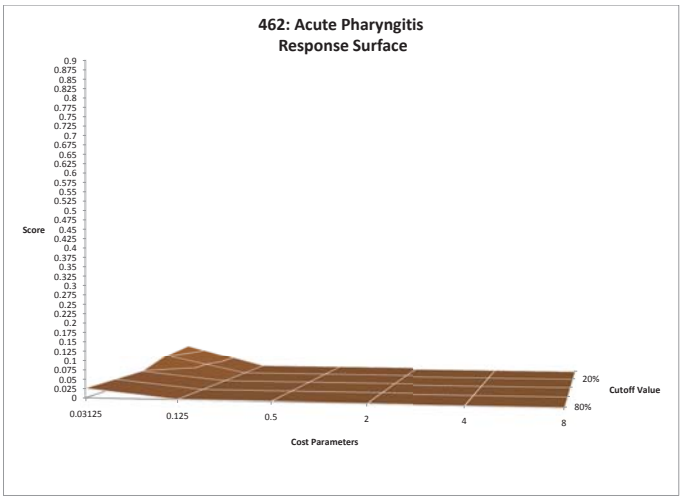


Figure B.6: The F-score response surface of 462 created by the tuning of the cost parameter and the confidence cutoff for term frequency-inverse document frequency weighting

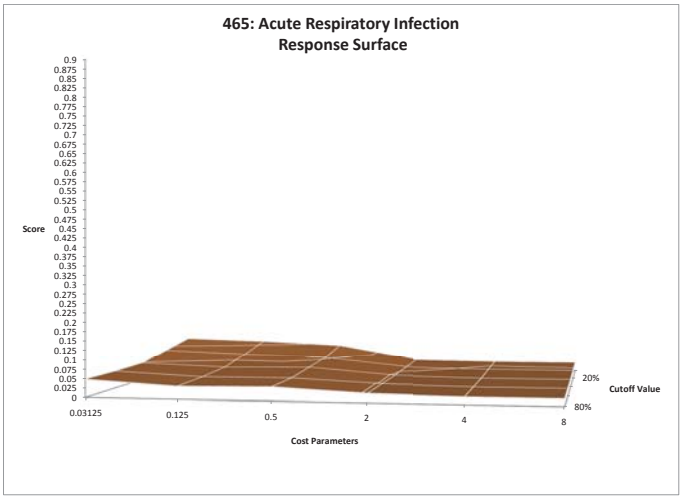


Figure B.7: The F-score response surface of 465 created by the tuning of the slack variable and the confidence cutoff for term frequency-inverse document frequency weighting

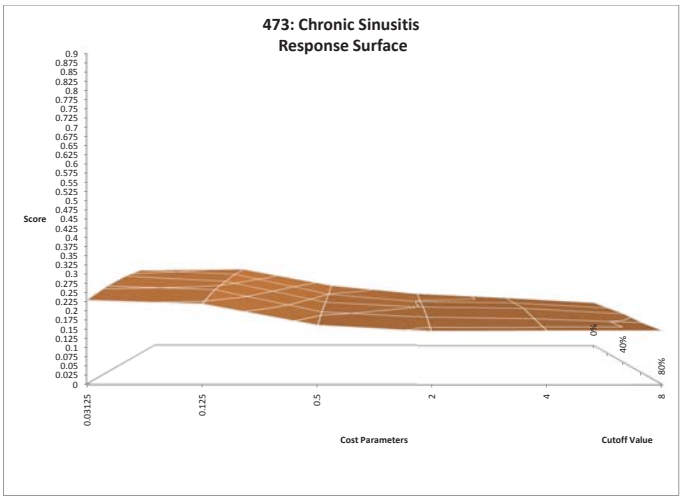


Figure B.8: The F-score response surface of 473 created by the tuning of the cost parameter and the confidence cutoff for term frequency-inverse document frequency weighting

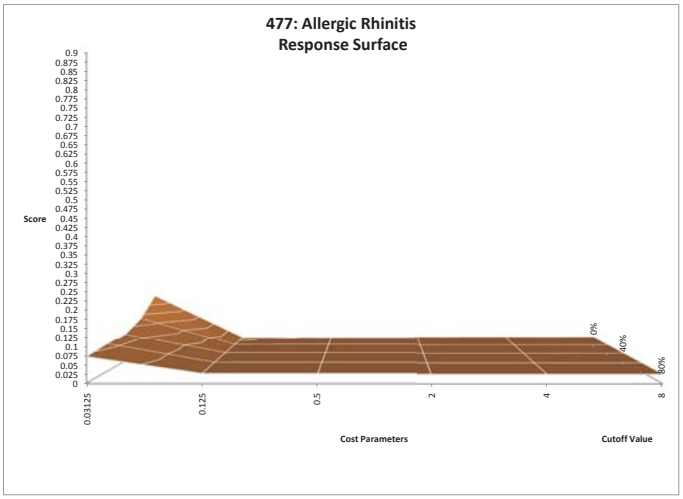


Figure B.9: The F-score response surface of 477 created by the tuning of the cost parameter and the confidence cutoff for term frequency-inverse document frequency weighting

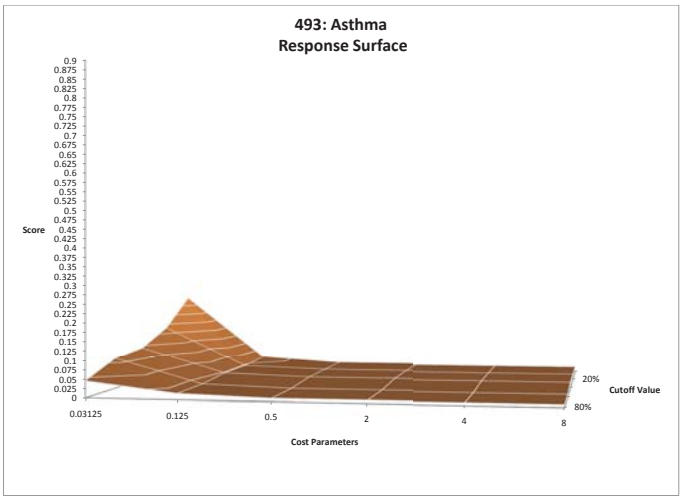


Figure B.10: The F-score response surface of 493 created by the tuning of the cost parameter and the confidence cutoff for term frequency-inverse document frequency weighting

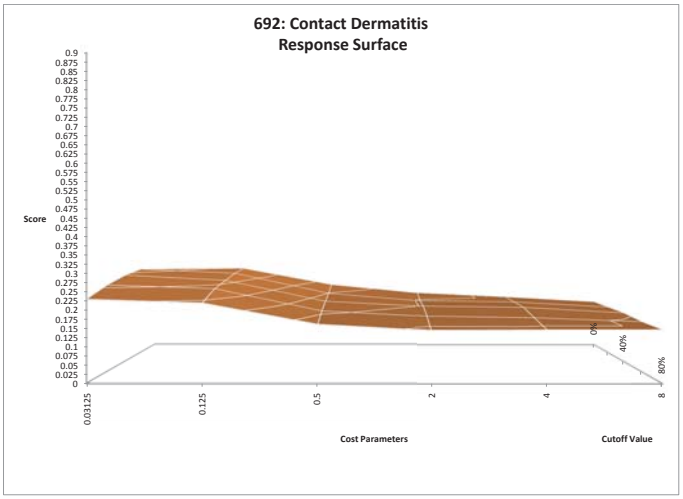


Figure B.11: The F-score response surface of 692 created by the tuning of the cost parameter and the confidence cutoff for term frequency-inverse document frequency weighting

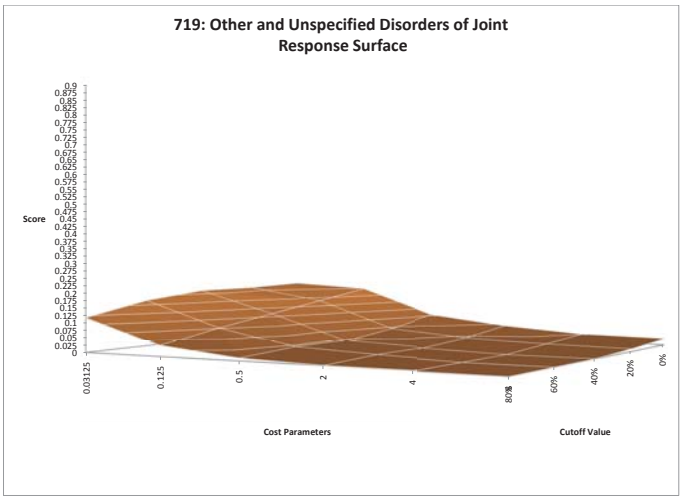


Figure B.12: The F-score response surface of 719 created by the tuning of the cost parameter and the confidence cutoff for term frequency-inverse document frequency weighting

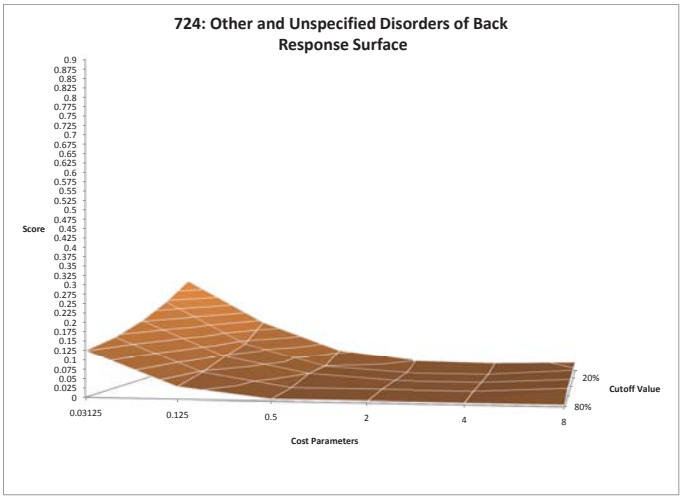


Figure B.13: The F-score response surface of 724 created by the tuning of the cost parameter and the confidence cutoff for term frequency-inverse document frequency weighting

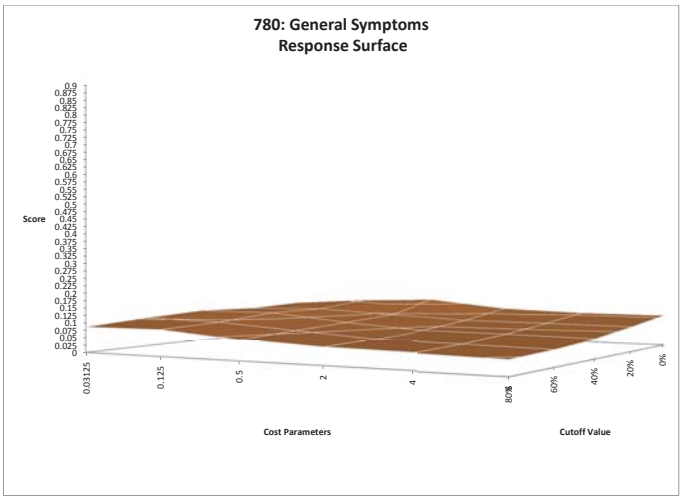


Figure B.14: The F-score response surface of 780 created by the tuning of the cost parameter and the confidence cutoff for term frequency-inverse document frequency weighting

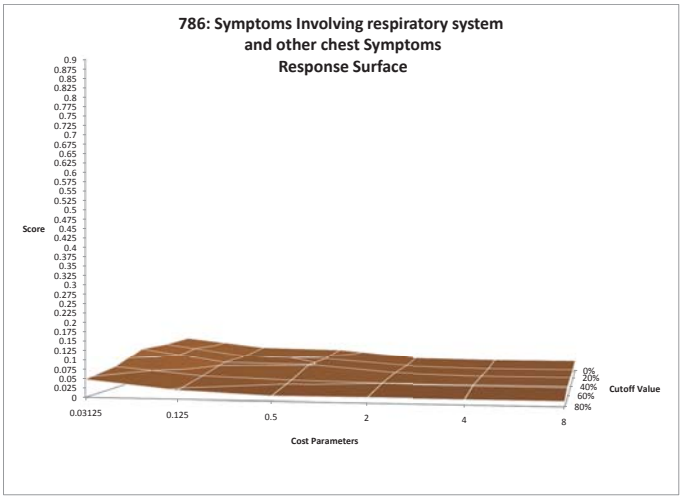


Figure B.15: The F-score response surface of 786 created by the tuning of the cost parameter and the confidence cutoff for term frequency-inverse document frequency weighting

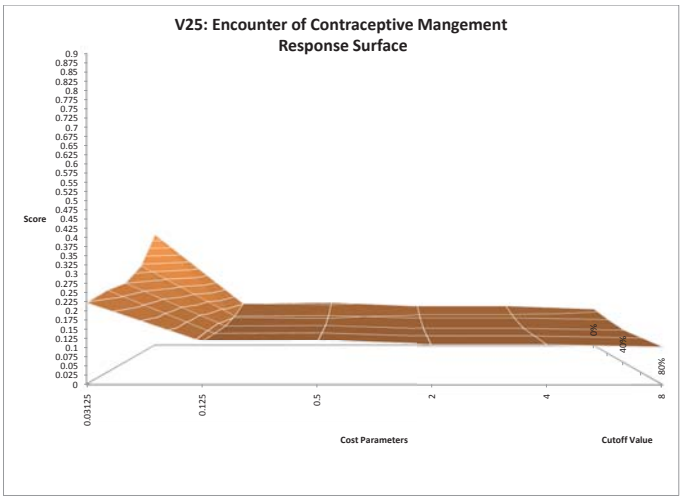


Figure B.16: The F-score response surface of V25 created by the tuning of the cost parameter and the confidence cutoff for term frequency-inverse document frequency weighting

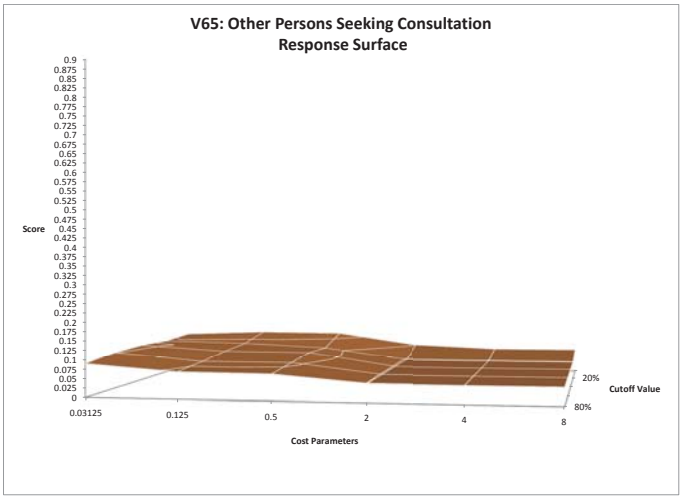


Figure B.17: The F-score response surface of V65 created by the tuning of the cost parameter and the confidence cutoff for term frequency-inverse document frequency weighting

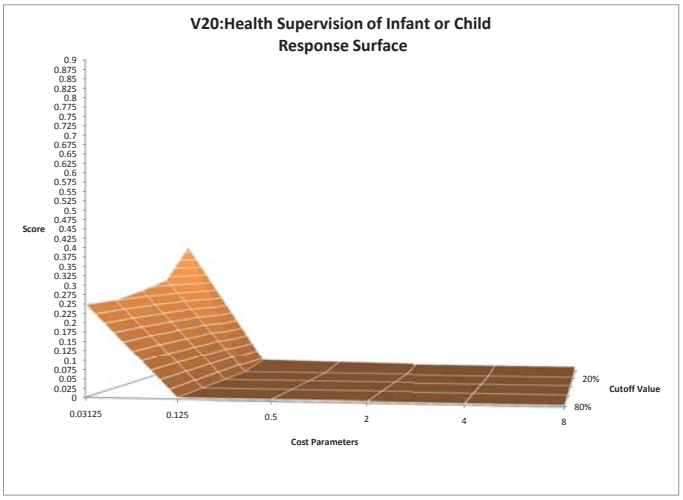


Figure B.18: The F-score response surface of V20 created by the tuning of the cost parameter and the confidence cutoff for term frequency-inverse document frequency weighting

APPENDIX C:
Additional Normalized Term Frequency-Inverse Document Frequency Response Surface

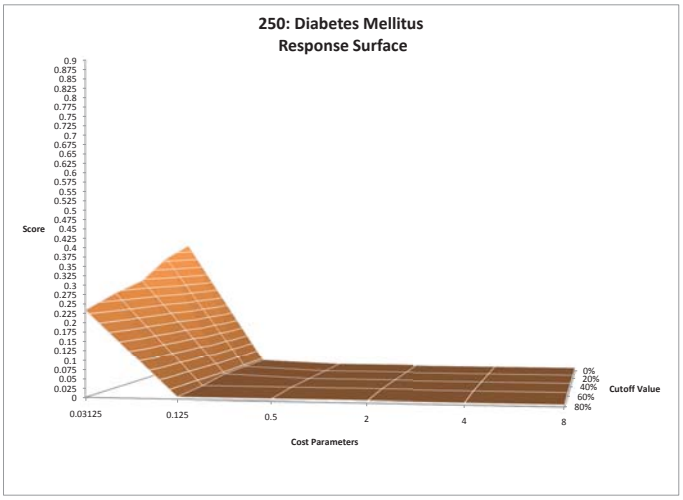


Figure C.1: The F-score response surface of 250 created by the tuning of the cost parameter and the confidence cutoff for normalized term frequency-inverse document frequency weighting

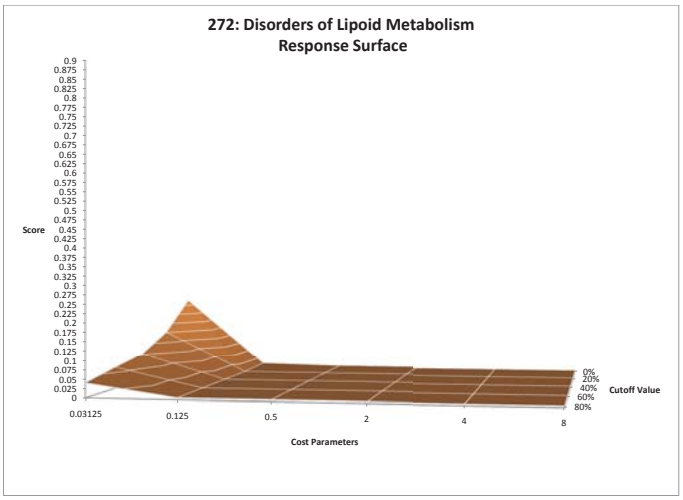


Figure C.2: The F-score response surface of 272 created by the tuning of the cost parameter and the confidence cutoff for normalized term frequency-inverse document frequency weighting

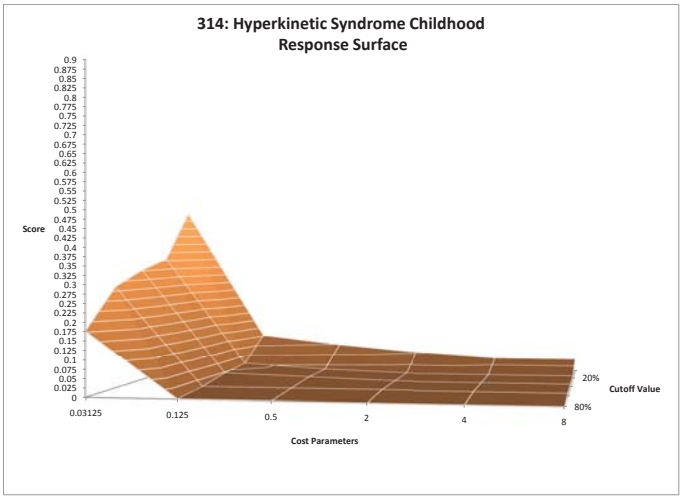


Figure C.3: The F-score response surface of 314 created by the tuning of the cost parameter and the confidence cutoff for normalized term frequency-inverse document frequency weighting

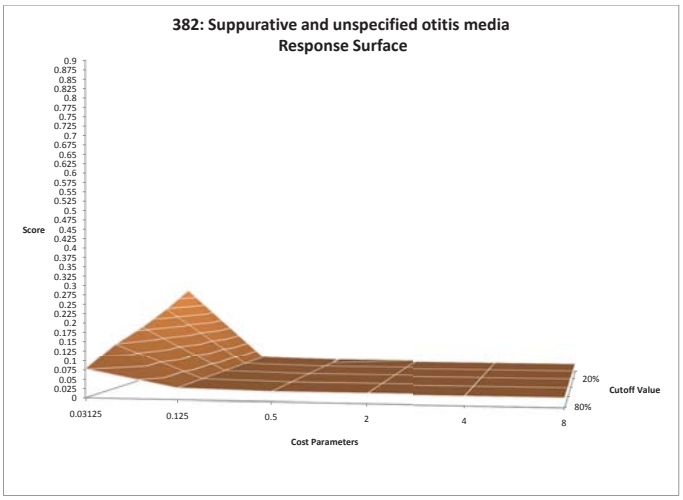


Figure C.4: The F-score response surface of 382 created by the tuning of the cost parameter and the confidence cutoff for normalized term frequency-inverse document frequency weighting

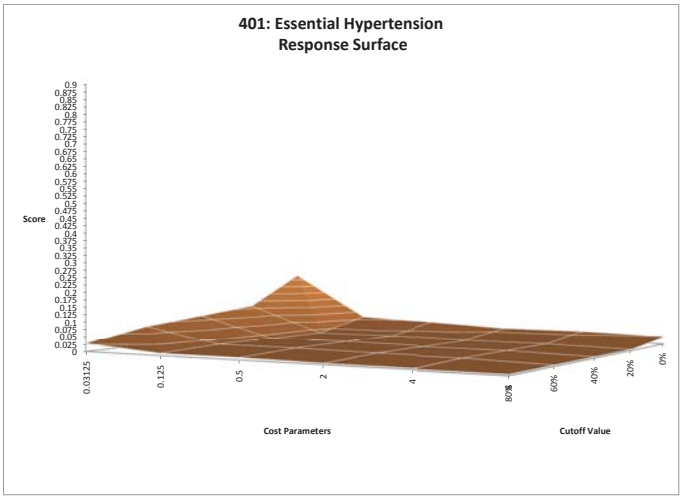


Figure C.5: The F-score response surface of 401 created by the tuning of the cost parameter and the confidence cutoff for normalized term frequency-inverse document frequency weighting

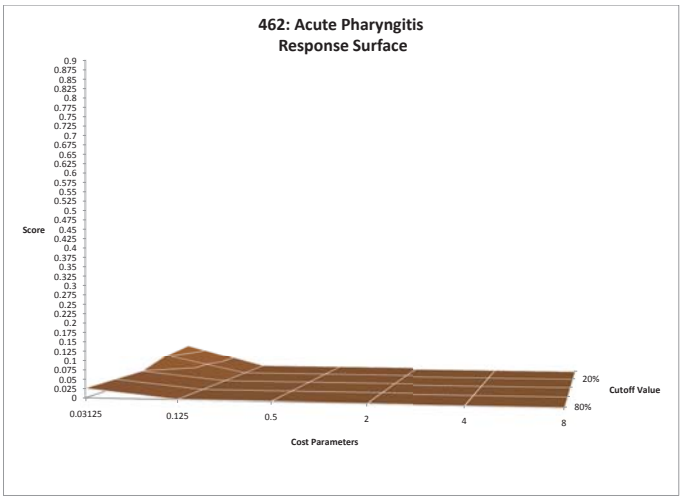


Figure C.6: The F-score response surface of 462 created by the tuning of the cost parameter and the confidence cutoff for normalized term frequency-inverse document frequency weighting

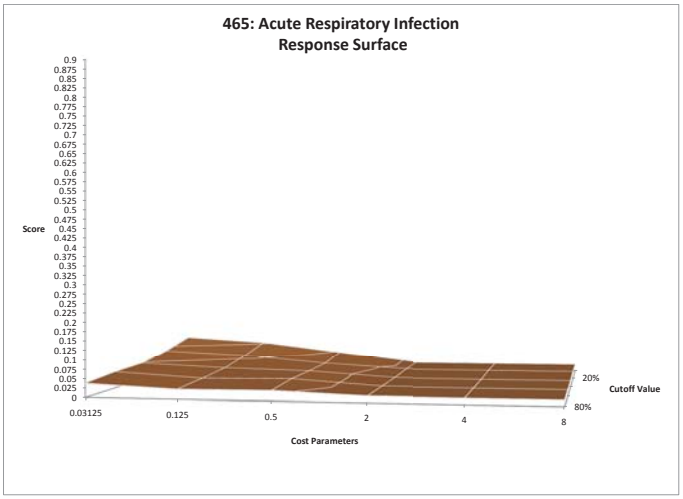


Figure C.7: The F-score response surface of 465 created by the tuning of the cost parameter and the confidence cutoff for normalized term frequency-inverse document frequency weighting

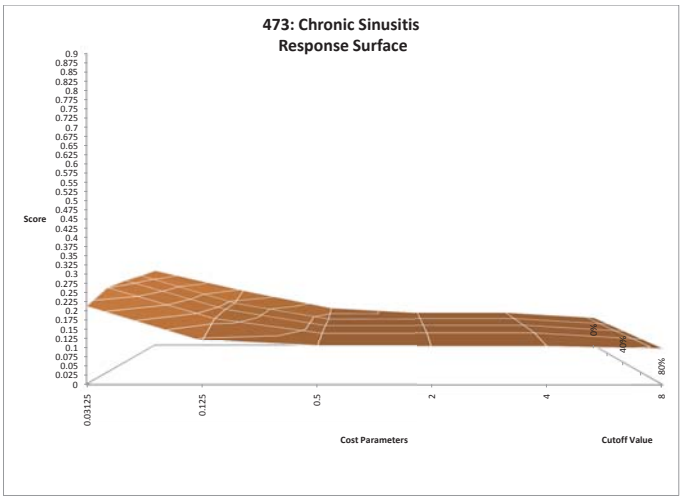


Figure C.8: The F-score response surface of 473 created by the tuning of the cost parameter and the confidence cutoff for normalized term frequency-inverse document frequency weighting

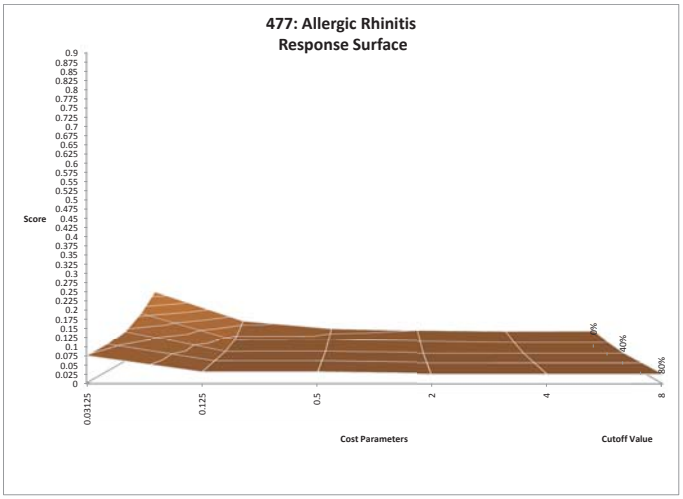


Figure C.9: The F-score response surface of 477 created by the tuning of the cost parameter and the confidence cutoff for normalized term frequency-inverse document frequency weighting

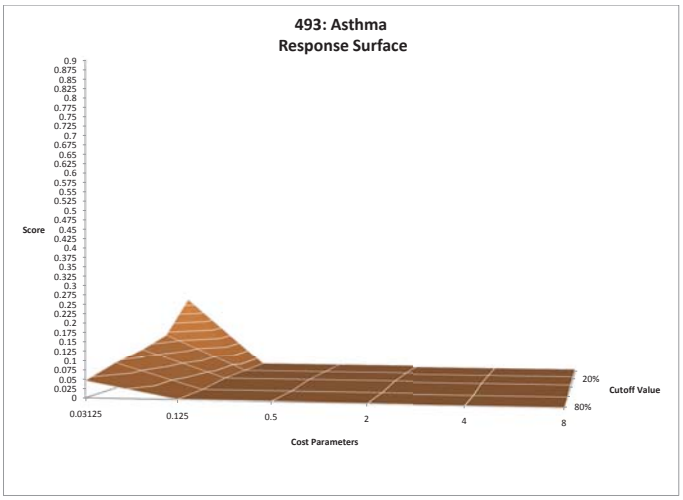


Figure C.10: The F-score response surface of 493 created by the tuning of the cost parameter and the confidence cutoff for normalized term frequency-inverse document frequency weighting

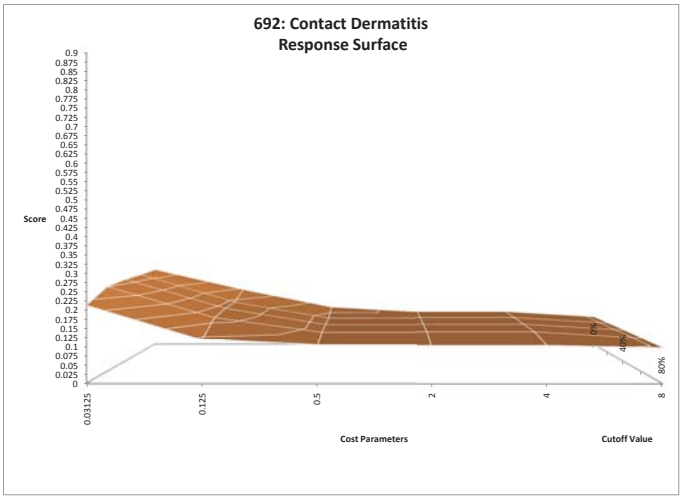


Figure C.11: The F-score response surface of 692 created by the tuning of the cost parameter and the confidence cutoff for normalized term frequency-inverse document frequency weighting

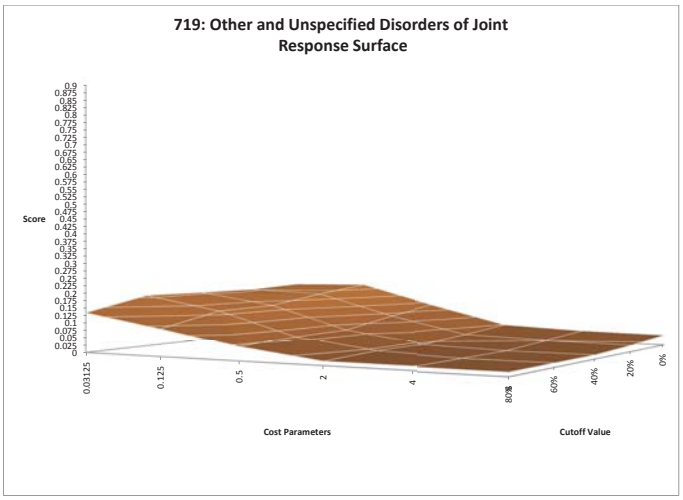


Figure C.12: The F-score response surface of 719 created by the tuning of the cost parameter and the confidence cutoff for normalized term frequency-inverse document frequency weighting

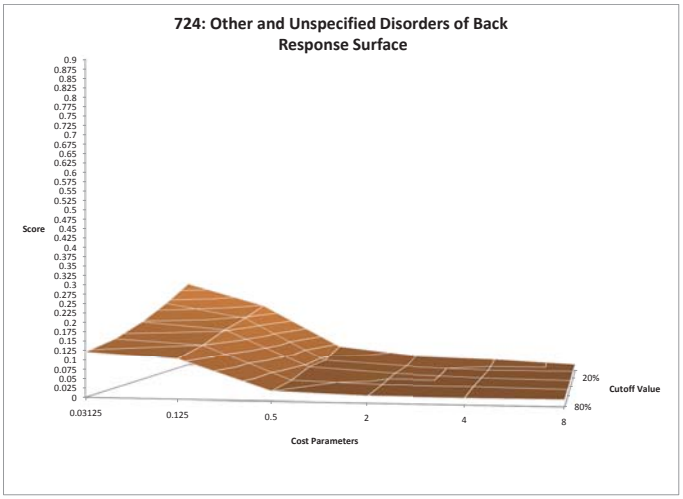


Figure C.13: The F-score response surface of 724 created by the tuning of the cost parameter and the confidence cutoff for normalized term frequency-inverse document frequency weighting

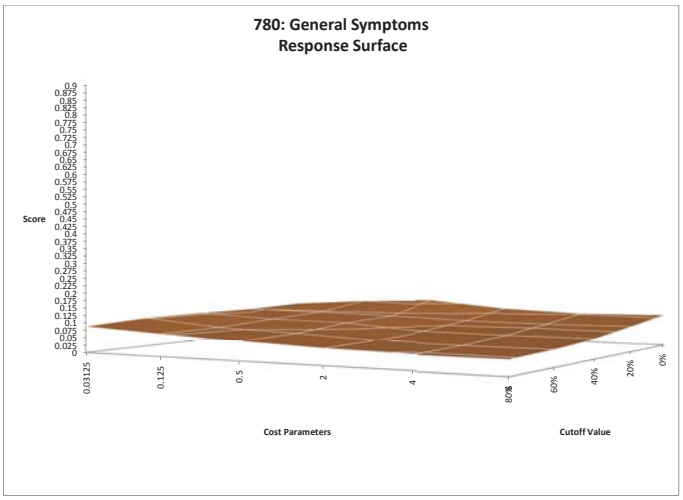


Figure C.14: The F-score response surface of 780 created by the tuning of the cost parameter and the confidence cutoff for normalized term frequency-inverse document frequency weighting

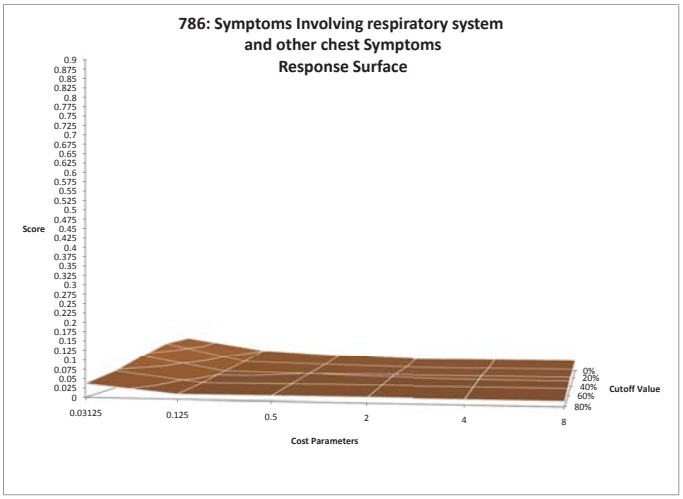


Figure C.15: The F-score response surface of 250 created by the tuning of the cost parameter and the confidence cutoff for normalized term frequency-inverse document frequency weighting

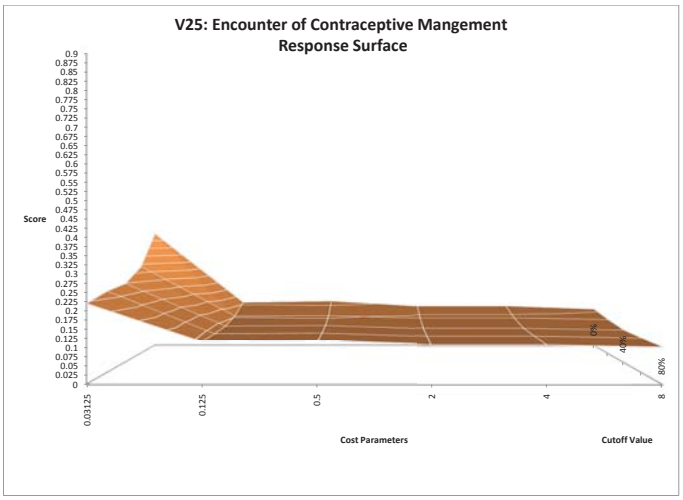


Figure C.16: The F-score response surface of V25 created by the tuning of the cost parameter and the confidence cutoff for normalized term frequency-inverse document frequency weighting

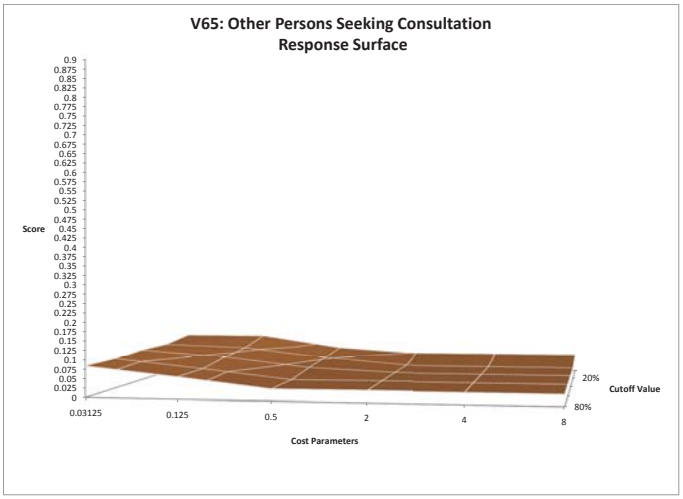


Figure C.17: The F-score response surface of V65 created by the tuning of the cost parameter and the confidence cutoff for normalized term frequency-inverse document frequency weighting

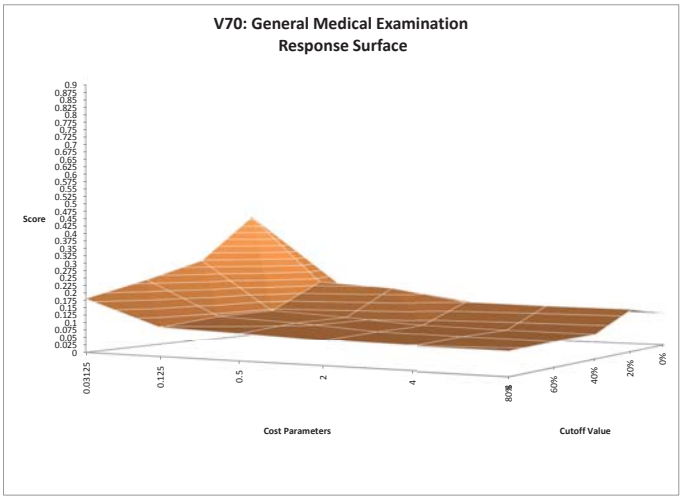


Figure C.18: The F-score response surface of V70 created by the tuning of the cost parameter and the confidence cutoff for normalized term frequency-inverse document frequency weighting

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia
2. Dudly Knox Library
   Naval Postgraduate School
   Monterey, California
3. Professor Lyn R. Whitaker
   Naval Postgraduate School
   Monterey, California
4. Professor Samuel L. Buttrey
   Naval Postgraduate School
   Monterey, California